The (un)willingness to reward cooperation and punish non-cooperation

Molenmaker, Welmer E.

Chapter 1

General introduction: Why study the willingness to sanction in social dilemmas?

■ Why study the willingness to sanction in social dilemmas?

The greatest challenge for all societies, regardless of how advanced they are, is to ensure and protect the collective welfare. This challenge arises from the fact that the interests of the collective do not necessarily coincide with the personal interests of the people belonging to that collective. Thus, on many occasions, people face a dilemma between furthering the collective interests or furthering their personal interests. Situations that revolve around this conflict of interests are called *social dilemmas* (Dawes, 1980). Consider, for instance, public goods and common resources to which people generally have unlimited access, such as medical care, public transport, national security, drinking water, energy, clean environments, etcetera. Despite the collective's interest to provide and/or preserve such entities, not all people may feel inclined to do so. After all, for an individual it is more profitable to consume from these entities without contributing to their provision or preservation (Olson, 1965; Samuelson, 1954). However, if too many people rather opt for their personal interests (i.e., non-cooperative choice behavior) than for the collective interests (i.e., cooperative choice behavior), public goods cannot be provided and common resources become depleted (Hardin, 1968). Thus, the pursuit of personal interests can have detrimental consequences for the collective welfare. How to prevent such collective failure? How to avoid the disastrous situation that unlimited access to public goods and common resources, as Garrett Hardin (1968) put it in his seminal paper *The Tragedy of the Commons*, eventually brings "ruin to all" (p. 1244)?

One of the most straightforward ways to protect the collective welfare is to make cooperative choice behavior more attractive and non-cooperative choice behavior less attractive (for overviews, see Kollock, 1998; Komorita & Parks, 1995; Messick & Brewer, 1983; Parks, Joireman, & Van Lange, 2013; Van Lange, Joireman, Parks, & Van Dijk, 2013; Weber, Kopelman, & Messick, 2004). To change the relative attractiveness of the options at hand, sanctions readily come to mind as effective solutions to accomplish this goal. Indeed, a recent meta-analysis including nearly 200 effect sizes demonstrated that both the use of positive sanctions (i.e., *rewards*) for cooperative choice behavior and negative sanctions (i.e., *punishments*) for non-cooperative choice behavior can effectively enhance cooperation (Balliet, Mulder, & Van Lange, 2011).[1] This issue of effectiveness has long dominated the scientific thinking on the use of sanctions in social dilemmas. Although it is an important insight that rewards and punishments are effective means to stimulate cooperative choice behavior, a critical question remained unanswered: Are people actually willing to sanction? This question is of critical importance, if only for the obvious reason that someone should first be willing to administer rewards and punishments before they can actually show their effects. For long, hardly any research dealt with this question (for an overview, see Van Dijk, Molenmaker, & De Kwaadsteniet, 2015). To shed more light on this neglected topic, the central theme of this dissertation is the (un)willingness to reward cooperation and punish non-cooperation.

---

[1] *Sanction* is the standard term to refer to both punishment (negative sanction) and reward (positive sanction) in various scientific domains (see e.g., Baldwin, 1971; De Kwaadsteniet, Rijkhoff, & Van Dijk, 2013; O'Reilly & Puffer, 1989; Van Lange, Rockenbach, & Yamagishi, 2014; Weiss & Sachs, 1991).

In the remainder of this first chapter I further introduce this central theme and argue why it is important to study the ingness to sanction in social dilemmas. I first elaborate on defining social dilemmas. This is followed by a brief review of the working of sanctions. Next, I address the ingness to sanction. That is, I explain why it is not self-evident that sanctions, even if they are considered effective, are administered in social dilemmas. Moreover, I address the value of having a better understanding of sanctioning. Finally, I outline the aims of this dissertation and, in doing so, discuss the methods used before I close with a short overview of the remaining chapters.

Defining social dilemmas

Why do social dilemmas pose such a key challenge for societies? Why is it often necessary to protect the collective welfare with the use of sanctions? Why does Hardin (1968) refer to social dilemmas as a 'tragedy'? At their core, social dilemmas are defined by two features: (1) for each individual it is more beneficial not to cooperate than to cooperate, but (2) all individuals are worse off if no one cooperates than if all cooperate (Dawes, 1980). Whereas the first feature prescribes that non-cooperation is the optimal strategy for the individual (i.e., it maximizes an individual's benefit at the lowest cost), the second feature prescribes that it is detrimental to the collective when everyone would follow this strategy. One can basically recognize these two features in almost any situation that involves interdependence among people (i.e., when people's actions affect one another). Social dilemmas take many forms. Two of the most important types are *public good dilemmas* and *common resource dilemmas* (Camerer, 2003; Dawes, 1980; Parks et al., 2013). I first elaborate on these two types of social dilemmas before I proceed to the question of why the conflict between personal interests and collective interests is considered a dilemma.

Public good dilemmas deal with situations in which goods and services can be realized for the public. Whereas people generally are free to use such public goods, the realization of public goods requires that people contribute to their provision (Olson, 1965; Samuelson, 1954). Consider, for instance, blood transfusions that are given to those in (medical) need, as long as there are enough blood donors to supply their provision. As the first feature of social dilemmas prescribes, the optimal strategy for the individual would be not to donate blood because contributing comes with a cost, while yielding no direct benefits. After all, people always get blood transfusions (if necessary) because it is not allowed to exclude anyone from blood transfusions, irrespective of whether or not they donated blood themselves. However, if too many people choose not to donate blood themselves, blood transfusions cannot be provided and all people will potentially be worse off, as is prescribed by the second feature of social dilemmas. Societies face many challenges that are essentially public good dilemmas, such as the provision of medical care, public transport, national security, education, and in a way one could even argue that world peace also is a public good dilemma.

In contrast, common resource dilemmas deal with situations in which common (scarce) resources can be consumed by a group of people. Whereas people generally are free to harvest from common resources, the preservation of common resources requires that people restrict their harvesting (Hardin, 1968; Ostrom, 1990). Take, for example, forests that provide wood,

as long as enough trees are left available to replenish the forest. Since excessive harvesting is more beneficial to the individual than restricted harvesting, the first feature of social dilemmas prescribes that the optimal strategy for the individual would be to cut down as many trees as possible. However, as the second feature of social dilemmas prescribes, if too many people choose to harvest excessively, forests deplete and all people will be worse off. After all, trees can only be harvested once and forests replenish very slowly. Many of the challenges that societies face are essentially common resource dilemmas, such as the preservation of drinking water, energy, clean environments, food, and on a global scale also our planet.

If non-cooperative choice behavior can have such disastrous consequences for the collective interests, why then is it so hard for people to constrain their pursuit of personal interests? Why is it not self-evident that people serve the collective interests and cooperate? Put differently, why are situations in which the collective interest collides with the personal interest considered *dilemmas*? To answer this question, it is important to note that it is generally assumed that people strive to maximize their own benefits at the lowest costs (the self-maximizing assumption). For example, in psychology the hedonic principle of seeking pleasure and avoiding pain is often considered basic to behavior (e.g., Kahneman, Diener, & Schwarz, 1999). All choices that people make can basically be traced back to the primary motive of maximizing pleasant feelings and minimizing painful feelings, which is perhaps best illustrated by the fact that both human and nonhuman species' choice behavior is modulated by brain regions associated with pleasure and pain (e.g., Leknes & Tracey, 2008). The self-maximizing assumption – which is besides psychology also leading in the other scientific disciplines that study people's choice behavior (i.e., anthropology, biology, economics, mathematics, political science, and sociology) – is essentially rooted in evolutionary theory.

As the evolutionary principle of adaptation stipulates, strategies that maximize an individual's benefit at the lowest cost in a particular environment (i.e., optimal strategies) are the most likely strategies to evolve and persist (Darwin, 1859/1962). After all, the individual's reproductive success – referred to as inclusive fitness (Hamilton, 1964) – would be lower with more costly strategies (i.e., suboptimal strategies). This process of natural selection thus causes optimal strategies to triumph over suboptimal strategies because they are more evolutionary adaptive. In evolutionary terms, optimal strategies are fitness maximizing strategies. To illustrate this principle of adaptation before applying it to social dilemmas, consider for instance the evolutionary advantage that great apes species (i.e., orangutans, gorillas, chimpanzees, bonobos, and humans) had over their ancestors more than 14 million years ago when they evolved the capacity to save and transport tools for gathering food (e.g., rocks to crack open nuts). While their ancestors had to depend on the tools available onsite, great apes species were able to bring the tools necessary to gather food (Mulcahy & Call, 2006). Gathering food was a less effortful endeavor (and thus more optimized), which gave great apes species an evolutionary advantage over their now extinct ancestors. This example demonstrates that optimal strategies are the most likely strategies to survive. As such, there is an evolutionary basis to assume that maximizing benefits at the lowest costs to the individual is hardwired into the genetic build of humans and any other species.

The idea that people choose optimal strategies over suboptimal strategies – often referred to as 'rational' choice behavior – is also one of the key assumptions in the mathematical analysis of interdependent situations, known as game theory (Von Neumann & Morgenstern, 1944). Game theory assumes that rational decision makers strive to maximize their own benefits at the lowest costs in interdependent situations, like chess players who play to win without losing any pieces. Based on this assumption, game theory provides a benchmark to predict the choice behavior of several rational decision makers in interdependent situations. This benchmark is known as a Nash equilibrium and reflects a rational decision maker's optimal strategy when all strategies of the other rational decision maker(s) are taken into account (Nash, 1950). The Nash equilibrium is called a dominant strategy because it is the optimal response to all strategies. This is often illustrated with a well-known dyadic game called the prisoner's dilemma (Poundstone, 1992), which also served as the blueprint for defining social dilemmas (Dawes, 1980). In the prisoner's dilemma, two individuals choose between furthering their personal interest (i.e., non-cooperative choice) or their joint interest (i.e., cooperative choice). As shown in Table 1.1, both individuals benefit more if they mutually cooperate (€3) than if they mutually not cooperate (€2). Yet, if only one individual cooperates, this cooperative individual benefits the least (€1), while the other non-cooperative individual benefits the most (€4). Since an individual always benefits more with the non-cooperative choice than with the cooperative choice – regardless of whether the other individual opts for the cooperate choice (€4 instead of €3) or the non-cooperative choice (€2 instead of €1) – the optimal and thus dominant strategy for each individual is not to cooperate.

From the above it becomes apparent that there are dominant strategies in interdependent situations like social dilemmas that optimize the personal interests. However, the prisoner's dilemma also nicely illustrates that Nash equilibria do not necessarily optimize the interests of the collective. Despite the fact that rational decision makers do not cooperate, collectively they are actually better off if they would cooperate. Behold the essence of social dilemmas: "individual rationality leads to collective irrationality" (Kollock, 1998, p. 183). It were these insights, although largely theoretical in nature, that made Hardin realize that collective disasters

Table 1.1. *Prisoner's dilemma*

		Individual 2	
		Cooperative choice	Non-cooperative choice
Individual 1	Cooperative choice	€ 3 / € 3	€ 4 / € 1
	Non-cooperative choice	€ 1 / € 4	€ 2 / € 2

would result from unrestricted freedom to pursue personal interests because this would eventually result in overpopulated societies that would overuse their scarce resources. In his seminal paper *The Tragedy of the Commons*, Hardin (1968) illustrates this point by describing a situation about herdsmen who are all allowed to let their herds graze on a common pasture. It is in each herdsman's personal interest (and thus individually rational) to let as many of their herds graze on the common pasture as possible. However, if all herdsmen do so, all grass will be eaten and destruction of the common pasture is inevitable (and thus collectively irrational). To protect the common pasture from destruction, herdsman should thus restrict their usage of the common pasture.

Since Hardin's seminal paper, a tremendous amount of empirical research has been done to examine the extent to which people indeed act as prototypical rational decision makers in social dilemmas (for comprehensive overviews, see e.g., Dawes, 1980; Gächter & Herrmann, 2009; Kollock, 1998; Komorita & Parks, 1995; Messick & Brewer, 1983; Parks et al., 2013; Pruitt & Kimmel, 1977; Van Lange, De Cremer, Van Dijk, & Van Vugt, 2007; Van Lange et al., 2013; Weber et al., 2004). Such empirical research has consistently shown that people are often not as 'rational' as game theory would predict, as they frequently cooperate with each other. The fact that cooperative choice behavior is quite ubiquitous in both human and nonhuman species thus suggests that cooperation has been evolutionarily adaptive in many situations. To give one example, when individuals are genetically related, cooperation could increase their reproductive success (for a clear overview of the evolutionary explanations why cooperation has evolved, see Barclay & Van Vugt, 2015). Whereas one can basically recognize a social dilemma in almost any situation that involves interdependence among people, there are circumstances that seem to disarm interdependent situations from its 'dilemma' (e.g., when people are close relatives). Yet, this cooperative course of action is very fragile and falls rapidly if people get the impression that others take advantage of their generosity (e.g., Fehr & Gächter, 2000; Hart, Bridgett, & Karau, 2001; Kameda, Tsukasaki, Hastie, & Berg, 2011; see Olson, 1965; Samuelson, 1954). Only a few non-cooperative 'bad apples' can already undermine the cooperative standards in a group (e.g., Kerr et al., 2009). So the challenges that social dilemmas pose are not so much rooted in the unwillingness of people to cooperate, it is the ease with which the balance may tilt towards non-cooperation that makes social dilemmas a real threat to the collective welfare of groups, organizations, and societies.

The working of rewards and punishments

Ever since political philosopher Thomas Hobbes published his influential book *Leviathan* (1651/1991), in which he argued that mutual cooperation can only be brought about by coercion, an often heard advice to 'solve' social dilemmas is to use punishments (e.g., Hardin, 1968; Olson, 1965; Ostrom, 1990). Punishments (but also rewards) are means that basically change the outcome structure of social dilemmas in such a way that cooperative choice behavior becomes more beneficial and non-cooperative choice behavior becomes less beneficial to the individual. However, because of its brutal Hobbesian appearance (Shinada & Yamagishi, 2007), punishment was often criticized and long ignored as proper solution to

social dilemmas (Crowe, 1969; Fox, 1985; Lynn & Oldenquist, 1986; Taylor, 1982). As a result, it took until the pioneering work of Toshio Yamagishi (1986) before punishment finally came into focus and sanctioning established itself as a major theme in social dilemma research (for an historical overview, see Shinada & Yamagishi, 2007). The purpose of the following paragraph is to provide a brief review of the working of sanctions.

Over the last decades, numerous experiments have consistently shown that punishments can enhance cooperation in social dilemmas (for overviews, see Balliet et al., 2011; Van Dijk et al., 2015; Van Lange et al., 2014). Yamagishi (1986, 1988) studied the effectiveness of punishment systems in repeated social dilemmas, and was the first to demonstrate that (after some time) the mere threat of punishment is enough to sustain high levels of cooperation. Whereas the studies by Yamagishi (1986, 1988) dealt with exogenous punishments (i.e., imposed by an external source), Ostrom, Walker, and Gardner (1992) revealed that people are also able to self-govern social dilemmas if they have the opportunity to punish each other (e.g., Fehr & Gächter, 2000, 2002; Gächter, Renner, & Sefton, 2008; Ostrom, Burger, Field, Norgaard, & Policansky, 1999). These kinds of endogenous punishments are often referred to as peer-to-peer punishment (Van Lange et al., 2014). Peer-to-peer punishment not only establishes high levels of cooperation in (small) groups, people also prefer groups in which they have the opportunity to punish each other over groups without any punishment opportunities (Gürerk, Irlenbusch, & Rockenbach, 2006). In their meta-analysis including 154 punishment effect sizes, Balliet et al. (2011) showed that punishment had a medium-sized (see also Cohen, 1988), positive effect on cooperation in social dilemmas (d = 0.70, 95% CI [0.60, 0.80]), especially if the punishment was costly to administer. According to Balliet et al., this indicates that the effectiveness of punishments seems to depend on whether people believe that these sanctions are administered with the intent to serve the collective interests.

One may conclude from the above that the challenges that social dilemmas pose can simply be solved by implementing punishments. However, punishment can also have negative effects. For example, Tenbrunsel and Messick (1999) showed that people generally perceive the decision they make in social dilemmas as an ethical decision (i.e., they consider the ethical aspects of their decision). However, when punishments are installed, this perception changes into a 'business' decision (i.e., they consider the costs and benefits of their decision). Consequently, people cooperate even less with weak punishments than they would do without punishments because they are more focused on the personal benefits of non-cooperation over cooperation (see Gneezy & Rustichini, 2000). Besides these negative effects of punishments on the willingness to cooperate, Mulder and colleagues (2006a) revealed that punishments also have negative effects on people's trust in others. The installation of punishments may communicate that the group does not consist of cooperative group members and that extrinsic means are needed to establish cooperation (see also Mulder, Van Dijk, & De Cremer, 2009; Mulder, Van Dijk, De Cremer, & Wilke, 2006b; Mulder, Van Dijk, Wilke, & De Cremer, 2005). Moreover, the effectiveness of punishment opportunities is undermined when people start to abuse them (e.g., Herrmann, Thöni, & Gächter, 2008; Rand & Nowak, 2011; for a review, see Sylwester, Herrmann, & Bryson, 2013). Non-cooperators may retaliate against

those who punished them (called counter-punishment), which usually is even more detrimental to the collective welfare than the mere presence of non-cooperators (Cinyabuguma, Page, & Putterman, 2006; Denant-Boemont, Masclet, & Noussair, 2007; Nikiforakis, 2008). In addition, cooperative choice behavior may be punished by non-cooperators (called antisocial punishment) when this cooperative choice behavior makes those who do not cooperate look bad (Parks & Stone, 2010). Against this background, one could say that punishments can solve social dilemmas, but only if those in control of punishments use them 'wisely'.

Whereas previous research has mainly focused on the effectiveness of punishments, rewards also change the outcome structure of social dilemmas in such a way that cooperative choice behavior becomes more beneficial and non-cooperative choice behavior becomes less beneficial to the individual. Yet, studies on the effectiveness of rewards are relatively scarce (for overviews, see Balliet et al., 2011; Van Dijk et al., 2015; Van Lange et al., 2014). This lack of attention might stem from the idea that cooperators already display desired choice behavior and it is therefore not necessary to reward them. However, rewarding cooperative choice behavior may also be an indirect punishment for those who are not rewarded. Moreover, the reward of cooperation may also send a signal about the desired behavior, also to those who do not cooperate (Van Dijk et al., 2015). As such, rewards may not only be a stimulant for those who already cooperate, it may also stimulate non-cooperators to change their choice behavior. The meta-analysis on 33 reward effect sizes by Balliet et al. (2011) showed that reward had a medium-sized (see also Cohen, 1988), positive effect on cooperation in social dilemmas (d = 0.51, 95% CI [0.31, 0.70]), especially if the reward was costly to administer. For example, McCusker and Carnevale (1995) adapted Yamagishi's (1986, 1988) experimental design and extended it with a reward system. Their results showed that rewards also sustain high levels of cooperation in social dilemmas. More recently, Rand and colleagues (2009) revealed that rewards even outperform punishments when both sanction means are available. However, this only seems to be the case in repeated social dilemmas in which people can build (positive) reputations (see also Rapoport & Au, 2001; Sefton, Shupp, & Walker, 2007; Sutter, Haigner, & Kocher, 2010; Walker & Halloran, 2004), which indicates that rewards are particularly effective in repeated interactions.

Just as for the implementation of punishments, rewards too can have negative effects. Research in other domains than social dilemmas, for example, showed that rewards may undermine autonomy and the intrinsic motivation to cooperate (e.g., Deci & Ryan, 2000; Ryan & Deci, 2000; for an overview, see Deci, Koestner, & Ryan, 1999). Furthermore, Chen, Pillutla, and Yao (2009) demonstrated that not only punishments but also rewards can have a negative effect on people's trust in others. The installation of extrinsic means – both punishments and rewards – may communicate that they are apparently needed because the group does not consist of intrinsically cooperative members (see Mulder et al., 2005). However, whereas people tend to evaluate (harsh) punishment of non-cooperation negatively (e.g., Atwater, Waldman, Carey, & Cartier, 2001; Eriksson, Andersson, & Strimling, 2015; Trevino, 1992; for a review, see Strimling & Eriksson, 2014), reward of cooperation is evaluated rather positively. For example, a study by Sutter et al. (2010) showed that people are more supportive of groups

that administer rewards than of groups that administer punishments. Furthermore, Kiyonari and Barclay (2008) revealed that those who administer rewards (but not punishments) are often rewarded in return (Milinski, Semmann, & Krambeck, 2002; Rand et al., 2009). Thus, the review presented above indicates that both rewards and punishments can be, and often are, effective tools to stimulate cooperative choice behavior, even though their use can also have negative effects. In addition, whereas the use of punishments is often accompanied with (some) public resistance, this seems less of an issue with the use of rewards. An emerging theme in social dilemma research therefore is whether people consider sanctioning the appropriate course of action. Central to this theme, and the prerequisite for any effect of sanctions, is the question whether people are actually willing to impose sanctions on others. In the following paragraph, I explain the value of addressing this important question.

The willingness to sanction

As a founding father of the contemporary research on sanctions in social dilemmas, Toshio Yamagishi (1986) was one of the first to stress the importance of the willingness to sanction. In social dilemmas, the administration of sanctions is often costly in terms of time, effort, and money. For example, the surveillance and monitoring of the Amazon rainforest to detect illegal logging is an effortful and difficult endeavor for local governments. Since the costs of sanctioning may exceed the benefits for an individual (Edney & Harper, 1978), Yamagishi recognized that the sole reliance on sanctions to solve social dilemmas may give rise to a new *second-order social dilemma* (Oliver, 1980; Yamagishi, 1986). Although all people in a group benefit when high levels of cooperation are established through sanctioning, there should first be enough people willing to incur the costs of administering sanctions. As the two features of social dilemmas prescribe (see Dawes, 1980), the optimal strategy for the individual would be not to participate in costly sanctioning, but it is detrimental to the collective when everyone would follow this strategy because cooperative choice behavior would not be enforced. Given the assumption that people choose optimal strategies over suboptimal strategies, there is a theoretical reason to believe that sanctioning is as unlikely as cooperative choice behavior. However, Yamagishi's (1986, 1988) early work on sanctioning demonstrate that people are willing to install punishments at their own expense if they value others' cooperation but expect or fear that others will not cooperate. More recently, Fehr and Gächter (2002) revealed that some people are even willing to costly punish non-cooperation when any direct gain for themselves is absent. To conclude, despite the fact that costly sanctioning is not self-evident, there are often people willing to use sanction opportunities (for an overview, see Van Dijk et al., 2015).

However, an important question that is often overlooked but needs to be addressed as well, is *why*? Why would people be willing to incur the costs of sanctioning in social dilemmas? Surprisingly enough, this fundamental question about the determinants of sanctioning has long remained unaddressed. Whereas it was generally assumed that people sanction to promote cooperative choice behavior and deter non-cooperative choice behavior (e.g., Yamagishi, 1986), more and more empirical evidence suggests that people do not

necessarily have the collective interest in mind when they reward cooperation and punish non-cooperation. For example, research on the motives underlying punishment indicates that retribution and not deterrence is the primary motive for punishment because people often want to give norm-violators their just deserts (e.g., Carlsmith, 2006; Carlsmith, Darley, & Robinson, 2002; Crockett, Özdemir, & Fehr, 2014; see also Mooijman, Van Dijk, Ellemers, & Van Dijk, 2015). In addition, studies have shown that people are even willing to sanction in single interactions without any repetition (i.e., one-shot games), where deterrence cannot play a role (Bone & Raihani, 2015; Gächter & Herrmann, 2009). These findings fit with the notion that the emotions that people experience in response to others' choice behavior typically fuel their willingness to incur the costs of sanctioning (e.g., De Kwaadsteniet et al., 2013; De Quervain et al., 2004; Fehr & Fischbacher, 2004; Fehr & Gächter, 2002; Nelissen & Zeelenberg, 2009; Pillutla & Murnighan, 1996; Seip, Van Dijk, & Rotteveel, 2014; Wang, Galinsky, & Murnighan, 2009). Taken together, it is too simplistic to assume that people use sanctions just to enforce cooperation.

I argue that there are several reasons why it is important to broaden the focus in social dilemma research to the determinants of the willingness to reward cooperation and to punish non-cooperation. First and foremost, this is important for our theoretical understanding of the psychological processes involved in the use of sanctions in social dilemmas. As mentioned above, emotions are often identified as a proximate mechanism underlying sanctioning (e.g., De Kwaadsteniet et al., 2013; Nelissen & Zeelenberg, 2009; Pillutla & Murnighan, 1996; Seip et al., 2014). However, other psychological processes – which received far less attention in experimental research – may also play a role (for an overview, see Van Dijk et al., 2015). Most importantly, the underpinnings of the willingness to sanction are not necessarily determinants that *foster* sanctioning, but may also be determinants that *hamper* sanctioning. Research on the *do-no-harm principle* has, for instance, shown that people tend to be reluctant to inflict harm on others (e.g., Baron, 1995; Baron & Jurney, 1993; Spranca, Minsk, & Baron, 1991), which may also apply to the use of sanctions. That is, the reluctance to harm may tone down the willingness to punish non-cooperative choice behavior, but not the willingness to reward cooperative choice behavior, even if this would imply that non-cooperation remains unpunished. Examining not only the determinants of sanctioning, but also their boundary conditions, may thus provide a more comprehensive understanding of the psychological mechanisms underlying the (un)willingness to sanction.

Second, focusing on the determinants of the willingness to sanction is also relevant for our theoretical understanding of the evolutionary functions of sanctioning in social dilemmas. In other words, analyzing the psychological determinants of sanctioning (i.e., the proximate level) can provide fruitful insights for analyzing the evolutionary functions of sanctioning (i.e., the ultimate level; Barclay & Kiyonari, 2014; Tinbergen, 1968). The emergence and maintenance of cooperation norms in groups has, for example, been suggested as an explanation of why sanctioning is evolutionary adaptive (Boyd, Gintis, Bowles, & Richerson, 2003; Fehr, Fischbacher, & Gächter, 2002; Fehr & Henrich, 2003; Gintis, 2000; Gintis, Bowles, Boyd, & Fehr, 2003; Henrich et al., 2010; Henrich et al., 2006).

This proposition about the evolutionary function of sanctioning has, however, been challenged by recent studies showing that the effectiveness of sanctioning is undermined when those who are being punished start to retaliate for the punishments they received (Cinyabuguma et al., 2006; Denant-Boemont et al., 2007; Nikiforakis, 2008; Rand, Armao, Nakamaru, & Ohtsuki, 2010). Consequently, alternative hypotheses about the evolutionary functions of sanctioning have been proposed that better fit these recent findings about the use of sanctions (e.g., Krasnow, Cosmides, Pedersen, & Tooby, 2012; Krasnow, Delton, Cosmides, & Tooby, 2016). The above illustrates how identifying determinants of sanctioning and their boundary conditions can steer the reasoning about the evolutionary forces that may or may not have selected for its design in new directions (see also Delton, Krasnow, Cosmides, & Tooby, 2011; Kenrick et al., 2009; Krasnow, Delton, Tooby, & Cosmides, 2013; Todd & Gigerenzer, 2007).

Third, besides the above theoretical considerations, there also is an important practical relevance to identify the determinants of the (un)willingness to sanction. Sanction opportunities may be implemented in real-life social dilemmas to make cooperation more attractive (by rewarding) and non-cooperation less attractive (by punishing). This requires, however, that the people in control of sanctions also use them for that purpose. For a clear understanding of how to implement sanction opportunities in real world social dilemmas, one should understand the conditions under which people actually use sanctions to promote cooperative choice behavior and deter non-cooperative choice behavior. Only then, one is able to organize sanction opportunities in such a way that they can be an effective solution to real-life social dilemmas. Thus, when it comes to solving social dilemmas in real-life, one first needs to understand the determinants of the (un)willingness to sanction before one can implement effective sanction opportunities. Research on the use of sanction may thus provide useful insights about the conditions that enable groups, organizations, and societies to ensure and protect the collective welfare.

■ The present research

The aim of the present dissertation is to identify determinants of the (un)willingness to reward cooperation and punish non-cooperation. Moreover, this research explores the psychological processes underlying sanctioning decisions. In doing so, I take the social psychological standpoint that an individual's choice behavior not only results from personal determinants, but also from situational determinants, as well as the interaction between both types of determinants (Lewin, 1951; Ross & Nisbett, 1991). Adopting this perspective is worthwhile since the use of sanctions in social dilemmas has been approached (almost) exclusively from an evolutionary perspective. As a consequence, prior research has mainly dealt with the question whether evolution selected for a disposition to promote cooperation – which would be a personal determinant of sanctioning – and its underlying mechanisms (Boyd et al., 2003; Fehr et al., 2002; Fehr & Henrich, 2003; Gächter & Herrmann, 2009; Gintis, 2000; Gintis et al., 2003; Henrich et al., 2010; Henrich et al., 2006). Yet, hardly any social dilemma research on the willingness to sanction dealt with the question of how the characteristics of

the situation affect the individual, which does in fact also teach us more about the motives that may underpin the (un)willingness to sanction. To fill this gap in the literature, the present work focuses on situational determinants of the (un)willingness to reward cooperation and punish non-cooperation.

Another aim of this dissertation is to examine whether people are as willing to punish non-cooperation as they are willing to reward cooperation. In the present dissertation, I argue and demonstrate that the willingness to punish differs markedly from the willingness to reward. In the context of social dilemmas, non-cooperative choice behavior tends to signal that the collective welfare may be jeopardized, while cooperative choice behavior tends to signal that things are going well. From this perspective, social dilemmas particularly call for deterrence of non-cooperation. Since non-cooperation is deterred more directly by punishing non-cooperation than by rewarding cooperation, one could argue that people should be more willing to punish non-cooperation than to reward cooperation. However, the use of punishments – in contrast with the use of rewards – implies that one directly inflicts harm on another person. Thus, punishment seems to come with the psychological 'cost' of harming others, whereas reward seems to come with the psychological 'benefit' of favoring others. As such, it may very well be that people actually prefer rewarding cooperation over punishing non-cooperation. The present work empirically tests this proposition. In doing so, I investigate whether the type of sanction people have at their disposal – either reward or punishment – is a primary determinant of their (un)willingness to sanction in social dilemmas.

Research approach

To set up the experiments presented in this dissertation, I used (1) economic games as experimental paradigm for social dilemmas and (2) financial sanctions as a measure of the willingness to sanction. With economic games (like the prisoner's dilemma described earlier), one creates rather straight-forward social decision making situations, which has a couple of key advantages. First, economic games capture the essence of social dilemmas (i.e., a conflict between personal and collective interests) without having to provide a context that has unintended connotations. Experimental research on economic games may thus generate fundamental insights about social decision making. Second, and related to the previous point, economic games are suited for an interdisciplinary approach, which is illustrated by the variety of scientific disciplines that use economic games to study social dilemmas, including anthropology, (evolutionary) biology, economics, mathematics, political science, psychology, and sociology. Third, economic games allow for direct comparisons between different types of social dilemmas because the games can be tailored in such a way that their payoff structures are identical. Fourth and finally, despite the simplicity in terms of structure, economic games are versatile in terms of the emotions, cognitions, and motives they may activate.

Furthermore, although sanctioning may manifest itself in many forms (from verbal compliments and reprimands to the infliction of physical pain), there are several reasons why financial sanctions – both bonuses and fines – provide a suitable approach to measure the willingness to sanction. First, financial sanctions are easy to implement in economic games

because both are numerical. Second, financial sanctions enable one to disentangle the choice to sanction from the size of sanctions. That is, the (un)willingness to sanction may not only be reflected by whether or not people engage in sanctioning, but also by the strength or size of the sanctions they administer. Third, and related to the previous points, financial sanctions allow for direct comparisons between the willingness to reward and the willingness to punish because the cost–consequence–ratio of both sanction means can be kept identical. Fourth and finally, financial sanctions can be presented without using the terms 'reward', 'bonus', 'punishment' or 'fine', which may have unintended connotations. To conclude, economic games and financial sanctions are the ideal methods to study the willingness to reward cooperation and punish non-cooperation in a laboratory setting.

Overview of the following chapters

The central theme of the present dissertation is the (un)willingness to reward cooperation and punish non-cooperation. In this first chapter, I introduced this central theme and argued why it is of critical importance to study the willingness to sanction in social dilemmas. The following three empirical chapters report the results of a series of experiments that I conducted to identify determinants of the (un)willingness to sanction. To identify sanction type (*Reward* versus *Punishment*) as a primary determinant, all empirical chapters (i.e., Chapters 2-4) revolve around the question how willing people are to reward cooperation and to punish non-cooperation. In the first two empirical chapters (i.e., Chapters 2 and 3), I argue that people generally prefer rewarding cooperation over punishing non-cooperation. In addition, each empirical chapter focuses on another situational factor, respectively *what* kinds of (non-)cooperative choice behavior people face (see Chapter 2), *how* they can sanction (see Chapter 3), and *when* they can sanction (see Chapter 4). In the present dissertation, I thus investigate whether the 'what', the 'how', and the 'when' are determinants of the (un) willingness to sanction. Note that the empirical chapters are based on separate articles that have either been published (Chapters 2 and 3) or are still under review for publication (Chapter 4). Although this makes it possible to read each chapter separately and in any order, it also creates some (minor) overlap between their opening paragraphs.

Chapter 2. This chapter – about what kinds of (non-)cooperative choice behavior people face – focuses on how the willingness to reward and punish depends on whether people face a public good dilemma or a common resource dilemma (Molenmaker, De Kwaadsteniet, & Van Dijk, 2014). I argue that people are less willing to punish non-cooperative choice behavior than to reward cooperative choice behavior. More importantly, however, I argue that the willingness to sanction is not only determined by the type of sanction people have at their disposal but is also moderated by the type of social dilemma they face. Specifically, I propose that people are less willing to punish and more willing to reward in public good dilemmas than in common resource dilemmas. The key findings of this chapter are that people punish less often and to a lesser extent than they reward, and that this general preference for rewarding cooperation over punishing non-cooperation is more pronounced in a public good dilemma than in a common resource dilemma (Experiments 2.1 and 2.2).

Chapter 3. This chapter – about how people can sanction – focuses on how personal responsibility has an impact on the (un)willingness to punish and reward (Molenmaker, De Kwaadsteniet, & Van Dijk, 2016). I argue that the general preference for the use of rewards over punishments is particularly pronounced when people feel personally responsible for administering the sanction. That is, I propose that people are reluctant to punish to the extent that they feel personally responsible for the harm done. Feelings of personal responsibility are manipulated by distinguishing between individual decision makers and joint decision makers (i.e., groups of people). The key findings of this chapter are that personal responsibility is an important determinant of the willingness to punish non-cooperation, but not of the willingness to reward cooperation (Experiments 3.1 and 3.2). In addition, this chapter reveals that feelings of personal responsibility have a self-restraining impact on the willingness to punish, regardless of people's external accountability (Experiment 3.3).

Chapter 4. This chapter – about when people can sanction – focuses on how the willingness to reward and punish is influenced by the timing of sanctioning decisions (Molenmaker, De Kwaadsteniet, & Van Dijk, 2016). I argue that people are less willing to sanction before than after the occurrence of others' choice behavior. In the decision environment beforehand others' actual choice behavior still has to take place in the future, whereas in the decision environment afterwards the choice behavior did actually take place in the past. I propose that people are less willing to sanction if the choice behavior has not occurred yet. The key findings of this chapter are that people are less willing to sanction choice behavior that may possibly occur in the future than choice behavior that did actually occur in the past. More specifically, people are less willing to reward and punish when they decide beforehand than when they decide afterwards (Experiments 4.1 and 4.2), regardless of whether they decide directly afterwards or after a time delay (Experiment 4.2).

Chapter 5. The final chapter provides an integration of all the previous chapters. Specifically, Chapter 5 contains a summary and discussion of the findings presented in this dissertation. In doing so, I report the results of two meta-analyses, one on the choice to sanction and one on the sanction size, that I performed on data reported in the empirical chapters of this dissertation and from experiments not included in these chapters (see Appendices A and B). Moreover, general implications of these findings are discussed and directions for future research are presented.