



Undermining trust and cooperation: The paradox of sanctioning systems in social dilemmas [☆]

Laetitia B. Mulder ^{a,*}, Eric van Dijk ^a, David De Cremer ^b, Henk A.M. Wilke ^a

^a *Department of Social and Organizational Psychology, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands*

^b *Department of Economic and Social Psychology, Tilburg University, The Netherlands*

Received 9 October 2003; revised 20 September 2004

Available online 27 April 2005

Abstract

Sanctioning systems in social dilemmas are often meant to increase trust in others and to increase cooperation. We argue, however, that sanctioning systems may also give people the idea that others act in their own self-interest and undermine the belief that others are internally motivated to cooperate. We developed the “Removing The Sanction” paradigm and a new trust manipulation, and showed in three experiments that when there is a sanction on defection, trust in others being internally motivated to cooperate is undermined: Participants who had experienced the presence of a sanctioning system trusted fellow group members less than participants who had not. In a similar vein, the sanction undermined cooperation when trust was initially high. The implications of these paradoxical findings are discussed.

© 2005 Elsevier Inc. All rights reserved.

Keywords: Social dilemmas; Sanctions; Trust; Cooperation

Introduction

Should I separate glass from my other garbage or should I throw all the garbage in the same dustbin? Should I fill in my tax form honestly or conceal certain incomes? Should I make full effort in my team assignment or free ride on my team members' efforts? Many of our daily life choices involve a conflict between our personal interests and collective interests. These situations are often referred to as “social dilemmas” (Dawes, 1980; Kopelman, Weber, & Messick, 2002; Liebrand, Messick, & Wilke, 1992; Messick & Brewer, 1983). In social dilemmas it is more profitable for individual group members to further their personal interests (defection) than further group interests (cooperation), whereas each individ-

ual obtains lower outcomes when everybody acts in their personal interests than when everybody acts in group interests.

One way to make people cooperate in social dilemmas is by sanctioning defection. Earlier research shows that such a sanctioning system can successfully increase cooperation (Caldwell, 1976; Eek, Loukopoulos, Fujii, & Gärling, 2002; Fehr & Gächter, 2002; McCusker & Carnevale, 1995; Van Vugt & De Cremer, 1999; Wit & Wilke, 1990; Yamagishi, 1986, 1992). The most straightforward explanation of the cooperation-increasing effect of a sanctioning system is that it changes the pay-off structure of the social dilemma: The attractiveness of defection is decreased and, therefore, cooperation becomes a relatively more attractive choice option.

However, with regard to sanctioning systems in social dilemmas research has more and more focused on the *negative* effects of sanctions. For example, Tenbrunsel and Messick (1999) showed that a sanctioning system can have negative effects on the perceived decisional

[☆] The authors thank Floor Rink and Kate Hudson for their comments.

* Corresponding author. Fax: +31 71 5273619.

E-mail address: lbulder@fsw.leidenuniv.nl (L.B. Mulder).

frame used in a social dilemma. In their research, people regarded the decision in a social dilemma situation as an ethical one, but as soon as a sanctioning system was installed, people regarded the decision as a business-related decision. In this way, the presence of a sanctioning system changed people's motives from ethical motives to more calculative motives. Similar ideas have been put forward in economics by Fehr and Falk (2001) and Frey and colleagues (Frey, 1993, 2000; Frey & Oberholzer-Gee, 1997) who argue that a financial incentive for moral behavior undermines the moral underpinning of that behavior by transforming norm violations into a kind of market transactions (see also Gneezy & Rustichini, 2000).

The negative effects of sanctions discussed above concern personal motives of individuals: Because of the sanction individuals regard their decision as a market-transaction, become more calculative and forget to see the moral aspects of the decision. So the sanction is argued to affect cognitions *within* individuals. Social dilemmas, however, are interdependent relations in which people's outcomes are not only affected by their own decisions, but also by the decisions of other people. Exactly in these situations, we argue it is important to focus on processes *between* individuals. After all, in social dilemmas, people's decisions are not only influenced by their *own* motivations to cooperate, but also for a great deal by the extent to which they expect *other* people to cooperate. In other words, *trust* that others will cooperate¹ is an important determinant of cooperation (e.g., Bruins, Liebrand, & Wilke, 1989; Coombs, 1973; De Cremer, Dewitte, & Snyder, 2001; De Cremer & Van Dijk, 2002; Kerr, 1983; Parks, Henager, & Scamahorn, 1996; Pruitt & Kimmel, 1977; Rapoport & Eshed Levy, 1989; Robbins, 1995; Schnake, 1991; Yamagishi, 1986). In this paper, we will focus on the relation between sanctioning systems and trust.

Sanctions and trust

A sanctioning system may not only increase cooperation by changing the pay-off, but also by increasing trust in other group members. In his *Leviathan*, Hobbes (1651/1909) argued in favor of coercive power and in his opinion, the main function of this coercive power is that it assures citizens that others will behave cooperatively. In line with this, Yamagishi (1986, 1992) showed that people tend to favor the introduction of a sanctioning system when there is little trust that other people will cooperate. Thus, a sanctioning system is supposed to

increase trust in other people involved in a social dilemma: Knowing that others will be punished if they defect results in more confidence that they will cooperate. This, together with the notion that trust positively relates to cooperation (Bruins et al., 1989; Coombs, 1973; De Cremer et al., 2001; Kerr, 1983; Parks et al., 1996; Rapoport & Eshed Levy, 1989; Robbins, 1995; Schnake, 1991; Yamagishi, 1986), suggests that trust created by a sanctioning system will increase cooperation. Thus, a sanctioning system may not only increase cooperation in a direct way, but also in an indirect way by increasing trust in others.

But what exactly does it mean when we say that a sanctioning system increases trust? Imagine yourself in a social dilemma situation in which a sanctioning system has been installed. Would it lead you to believe that others will cooperate because they genuinely care about the well being of the group? Or would it lead you to believe that others will cooperate simply because they wish to avoid the sanction placed on defection? More precise, trust can have various grounds (Blomqvist, 1997).

Whilst it may be true that a sanctioning system creates expectations that other people will cooperate, we seriously question whether it enhances the belief that others are internally motivated to cooperate. After all, the presence of a system more or less enforces cooperation rather than letting people choose cooperation voluntarily (Luhmann, 1988). Therefore, people might base their trust on the presence of the sanctioning system (cf. "institutional trust," Sitkin & Roth, 1993). In other words, they might believe that cooperation of others is externally motivated, that is, motivated by external incentives. This type of trust contrasts with trust based on the belief that others are internally motivated to cooperate, that is, the belief that others will choose behavior voluntarily, without trying to attain an external incentive (cf. the distinction between intrinsic and extrinsic motivation, Deci, 1971; Deci, Benware, & Landy, 1974; Deci, Koestner, & Ryan, 1999; Ryan & Deci, 2000). In this paper, we distinguish between these two bases of trust: between the trust in others being externally motivated to cooperate (because there is a coercive system) and the trust in others being internally motivated to cooperate.

We argue that a sanctioning system may increase people's trust in others being externally motivated to cooperate, but that it does not increase trust in others being internally motivated to cooperate. In fact, we argue that a sanctioning system can *decrease* trust that others are internally motivated to cooperate. Our reasoning for this is that a sanctioning system could serve as a message of distrust towards group members (Cialdini, 1996; Tenbrunsel, 1999). By installing a sanctioning system, an authority might implicitly communicate that there is reason to doubt that group members will cooperate "by themselves" (after all: Why else would there be a sanctioning system?). Additionally, although group members

¹ There are different definitions of trust in literature, differing from those highlighting the social and ethical facets of trust to those emphasizing the strategic and calculative dimension of trust (see for an overview, Kramer, 1999). What we mean by trust in this paper is the extent to which someone believes another person will cooperate.

might expect others to cooperate in the presence of a sanctioning system, they could attribute other people's cooperation to the presence of the sanctioning system rather than to these other people being internally motivated to cooperate. In this way, attributing cooperation to a sanctioning system can diminish perceived trust in others being internally motivated to cooperate. For these reasons people may conclude from the presence of a sanctioning system that their fellow group members pursue their own self-interest.

Findings from outside social dilemma research support our argument. First, Strickland's (1958) participants in the role of supervisor had to supervise two subordinates doing a dull task. The possibility for the supervisor to monitor and sanction was larger for one subordinate than for the other, so that there was a "frequently monitored subordinate" and a "scarcely monitored subordinate." At the end of the experimental session, the supervisor trusted the frequently monitored subordinate less to work hard than the scarcely monitored subordinate. Although he did not find statistical support for it, the reason for this result according to Strickland was that subjects attributed their frequently monitored subordinates' effort to external causes rather than internal. Second, De Dreu, Giebels, and Van de Vliert (1998) varied the extent to which negotiators could punish each other. Negotiators' opportunity to punish each other appeared to increase (threats of) punishments and led to decreased trust among negotiators.

It has to be noted that the studies of Strickland (1958) and De Dreu et al. (1998) concerned trust between "punishers" and "punishees." Strickland's study concerned the effect of punishing a person on trust in that punished person and De Dreu et al.'s study concerned the effect of (the threat of) being punished on trust in the punisher (see for a similar effect in the field of economics, Fehr & List, 2002). The effect of sanctions on trust in these studies may have been the direct cause of *the act of* (threatening with) sanctioning. We argue, however, that it is the sanctioning system *itself* that can undermine trust. In the contexts of social dilemmas, this means that a sanction imposed by a super ordinate authority can affect the trust levels between the *group members*. We argue that the mere presence of a sanctioning system (installed by someone outside the group) is enough to undermine trust among group members.

In sum, we argue that although a sanctioning system can increase trust in fellow group members, this trust will be primarily based on the presence of the sanctioning system (trust in others being externally motivated to cooperate). Moreover, trust in others being *externally* motivated to cooperate can replace trust in others being *internally* motivated to cooperate: Because the presence of the sanctioning system will give people the idea that fellow group members are self-interested, trust in others being internally motivated to cooperate will be undermined.

In this paper, we tested the negative effect of a sanctioning system on trust in others being internally motivated to cooperate. An issue when testing this is, however, that the undermining effect of a sanctioning system on trust in others being internally motivated to cooperate may not surface in the presence of that sanctioning system. Simply because a sanctioning system *increases* trust in others being externally motivated to cooperate, the overall trust level may not reveal that the sanctioning system has at the same time *decreased* trust that others are internally motivated to cooperate. This undermined trust that others are internally motivated to cooperate, however, may manifest itself in a subsequent situation in which a sanctioning system is absent. After all, in such a situation the reason to trust that others are externally motivated to cooperate is absent, making it necessary for people to base their trust purely on trust that others are internally motivated to cooperate. Therefore, we tested the effect of a sanctioning system on trust in others being internally motivated to cooperate using a newly developed method that we will refer to as the "Removing The Sanction" (RTS) paradigm.

In the RTS paradigm the level of trust is compared between people who have previously experienced a sanctioning system and people who have not. This situation is created by subsequently presenting participants two social dilemmas in which they have to decide to what extent they will cooperate. In the first phase of the experiment, a social dilemma is presented in which either a sanctioning system is present (sanction condition) or not (no-sanction condition). In the second phase of the experiment, the same kind of social dilemma situation is presented again. The only difference with the first phase is that in the second phase, the sanctioning system in the sanction condition is no longer present. As a result, in the second phase two groups of participants can be distinguished that differ in having experienced a sanctioning system in the previous phase (the sanction condition) or not (the no-sanction condition). It is in phase 2 that the undermining of trust that others are internally motivated to cooperate is expected to reveal itself in a lower trust level in the sanction condition compared to the no-sanction condition.

Because of a sanction's expected opposing effects (i.e., increasing trust in others being externally motivated to cooperate and decreasing trust in others being internally motivated to cooperate), we do not expect an overall effect of sanction on *phase 1* trust. Our primary interest with regard to trust, however, is in the second phase rather than the first. It is in the *second* phase where the undermining of trust in others being internally motivated to cooperate should come to light, because in the second phase the sanctioning system is not present (any more) and the level of trust cannot be based on its presence. Thus, phase 2 trust being lower in the sanction

condition than in the no-sanction condition would indicate that the sanctioning system has undermined trust in others being internally motivated to cooperate. Also, it would indicate that this undermined trust could manifest itself in a different situation in which a sanctioning system is absent.

As stated earlier, previous research has shown that the level of trust affects cooperation (Bruins et al., 1989; Coombs, 1973; De Cremer et al., 2001; Kerr, 1983; Parks et al., 1996; Rapoport & Eshed Levy, 1989; Robbins, 1995; Schnake, 1991; Yamagishi, 1986). With regard to the trust–cooperation relationship, we are also primarily interested in phase 2 effects. In phase 1 a sanction may probably increase cooperation, simply because defection is made a less attractive option. Moreover, in phase 1, overall trust both comprises trust in others being externally and trust in others being internally motivated to cooperate. All this may cause cooperation in phase 1 to be unrelated to people's overall trust level. In phase 2, however, the sanction as a direct reason to cooperate is gone. And if the previous sanction indeed appears to have undermined trust in other people's internal motivation to cooperate, we may expect that to affect the level of cooperation in phase 2 as well. Therefore, we expect trust and cooperation to be positively related in phase 2 and expect phase 2 cooperation to be lower in the sanction condition than in the no-sanction condition.

Study 1 provided a first test of our ideas. In the second and third studies, we further explored the relation between a sanctioning system and trust by including the level of initial trust as a factor in our research design.

Study 1

As a first test of our ideas, two conditions (sanction versus no-sanction) were compared. We expected a drop in trust and cooperation following removal of the sanctioning system. More important, in the sanction condition, we expected in the second phase of the experiment, trust and cooperation to drop below the level of the no-sanction condition. Thus, in the second phase we expected less trust and less cooperation in the sanction condition than in the no-sanction condition.

Method

Design and participants

For this experiment, 50 students (12 male and 38 female, mean age of 21.5 years) of the University of Amsterdam participated to fulfill a course requirement in the first year of their study. The experiment lasted 30 min and the participants were randomly assigned to the two conditions.

Procedure

Participants sat in experimental cubicles containing a table, a chair, and a computer. Communication with other participants was prohibited. The instructions for the experiment appeared on the computer screen.

Participants learned that they formed a group together with three other participants who were present at the same time in a similar cubicle. They were told that two situations (A and B) would be presented to them in which they could earn money. The money earned in the experiment would be paid to one of the groups participating in the experiment. To make sure participants regarded situations A and B as independent situations, we told them that the money earned in either situation A or B would be paid to this group. Which group would be paid and whether this would be the money earned in situation A or in situation B, would be determined randomly after the experiment.

Participants were told that in situation A (i.e., phase 1) group members each possessed 100 chips, each worth EUR 0.05. Group members could give any number of these chips to the group (i.e., cooperate) or keep them for themselves (i.e., defect). The total number of all chips given to the group by the four individual group members would be doubled and divided equally amongst the four group members. In this way, it was better for all group members if individual group members would donate chips to the group than if they kept chips for themselves. Individual group members, however, profited more by keeping chips for themselves than by donating chips to the group. After all, it is true that the chips a group member donated to the group would be doubled, but that group member would only receive a quarter of that. So, for each chip an individual donated to the group half a chip would return to that individual.

Participants in the sanction condition were told that the two group members who gave the least number of chips to the group would be fined EUR 5 (McCusker & Carnevale, 1995; Van Vugt & De Cremer, 1999). In the no-sanction condition, there was no mention of a sanctioning system.

Then, overall trust in group members was measured by asking participants whether they trusted the other group members as a whole to give chips to the group. This question was answered on a scale ranging from 1 (*absolutely not*) to 7 (*absolutely*). After this, participants were asked how many chips they decided to give to the group, which formed a measure of cooperation.

No feedback about others' decisions was given. After participants had made their decision in situation A, situation B (i.e., phase 2) was presented. Participants were told that in situation B they were asked to make the same kind of decision as in situation A and that again all group members possessed 100 chips of EUR 0.05, and all chips given to the group would be doubled and divided equally amongst the group members. In the sanction

condition, it was added that, in situation B, there would no longer be a sanctioning system. Participants in this condition were told that the absence of the sanctioning system was independent of their past decisions. Then, overall trust was measured with the same question as in situation A. After this, participants were asked how many chips they wanted to give to the group.

Finally, participants were thanked and fully debriefed. As indicated earlier, random payment of one group was made according to the choices its group members made in situation B.

Results

We first performed analyses on trust in other group members in phase 2 compared to phase 1, and on trust in other group members in phase 2. After this, to explore the effect on cooperation, we performed the same analyses on the number of chips that participants gave to the group.

Trust

To assess whether trust as a function of sanction had decreased in phase 2 compared with phase 1, we performed a 2 (sanction) \times 2 (phase) ANOVA with phase as a within-subjects factor. This analysis yielded a main effect of phase ($F[1,48]=11.69, p=.001$), qualified by a Sanction \times Phase interaction ($F[1,48]=13.10, p<.001$). The main effect of phase demonstrates that trust was lower in phase 2 ($M=4.18, SD=1.53$) than in phase 1 ($M=4.86, SD=1.37$). The Sanction \times Phase interaction (see upper part of Table 1) shows that there was only a decrease in trust in the sanction condition ($t[24]=3.79, p<.001$) and not in the no-sanction conditions ($t[24]=-0.27, ns$).

Then, we tested whether, in phase 2, trust in the sanction condition had decreased below the level of the no-sanction condition. In agreement with our hypothesizing, trust was lower in the sanction condition ($M=3.64, SD=1.52$) than in the no-sanction condition ($M=4.72, SD=1.37$), $t(48)=2.64, p=.01$.

Cooperation

To investigate whether cooperation had decreased in phase 2, compared with phase 1, as a function of sanc-

tion, we performed a 2 (sanction) \times 2 (phase) ANOVA on cooperation with phase as a within-subjects factor. This analysis yielded a marginal significant main effect of phase ($F[1,48]=3.95, p=.053$), qualified by a significant Sanction \times Phase interaction ($F[1,48]=9.59, p=.003$). The main effect of phase demonstrates that cooperation was lower in phase 2 ($M=53.76, SD=34.91$) than in phase 1 ($M=61.50, SD=31.93$). The marginal significance of this main effect significant may have been due to both the relatively high standard deviations and the fact that the main effect was dependent on sanction, as shown in the Sanction \times Phase interaction (see lower part of Table 1). This interaction shows that there was only a decrease in cooperation in the sanction condition ($t[24]=3.22, p=.004$) and not in the no-sanction condition ($t[24]=-0.90, ns$). Thus, cooperation decreased in phase 2 compared to phase 1, but this was only the case in the sanction condition and not in the no-sanction condition.

To investigate whether in phase 2 cooperation in the sanction condition had decreased below the level of the no-sanction condition, we also compared the mean level of phase 2 cooperation for both conditions. The level of phase 2 cooperation in the no-sanction condition was higher ($M=56.32, SD=32.65$) than in the sanction condition ($M=51.20, SD=37.53$). This difference was in the expected direction, but non-significant ($t[48]=0.52, ns$).

Additional analysis

So far, the results show a sanction effect on phase 2 trust, but not on cooperation. This might imply that cooperation was not related to trust in phase 2. A correlational analysis, however, did indicate that trust and cooperation in phase 2 significantly correlated ($r=.33, p=.02$). Thus, trust and cooperation in phase 2 were related.

Discussion

The results of Study 1 show that both trust and cooperation decrease when an existing sanctioning system is removed. As hypothesized, after removing the sanction, trust even decreased below the level of trust in the no-sanction condition. So, participants might have trusted fellow group members to cooperate as long as a sanctioning system was present. But this trust appeared to be conditional on the presence of the system. This process came to light when the sanctioning system was removed: Trust in fellow group members was then lower for participants who had experienced the presence of a sanctioning system in the previous phase than for those who had not experienced this. This finding indicates that the former presence of a sanctioning system undermined trust in others being internally motivated to cooperate. Although the level of cooperation also decreased after the sanctioning system was removed, the level of

Table 1
Trust and cooperation as functions of phase and sanction, Study 1

	Mean		SD	
	Phase 1	Phase 2	Phase 1	Phase 2
<i>Trust</i>				
No sanction	4.68 ^a	4.72 ^a	1.46	1.37
Sanction	5.04 ^a	3.64 ^b	1.27	1.52
<i>Cooperation</i>				
No sanction	52.00 ^a	56.32 ^a	32.98	32.65
Sanction	71.00 ^b	51.20 ^a	28.40	37.53

Note. For each row, different superscripts differ significantly ($p<.05$, paired-samples t test).

cooperation in the sanction condition did not significantly drop below the level of cooperation in the no-sanction condition.

The sanction thus undermined trust that others are internally motivated to cooperate, but did not undermine cooperation. Was this because trust was not related to cooperation? The moderate correlation between trust and cooperation in phase 2 does not support such an interpretation. The fact that the findings on trust did not perfectly mirror the findings on cooperation, however, implies that cooperation in Study 1 was not purely based on trust. In the next studies, we will come back on the relation between trust and cooperation in phase 2 and we will further address this in the General discussion.

Study 2

The way sanctions relate to trust may depend on the level of trust that people have in each other's cooperation in the first place. In certain social dilemma situations in which people have little trust that others will cooperate to begin with, introducing a sanctioning system is often the efficient thing to do (Yamagishi, 1986). Even though the trust that will arise from the installation of a sanctioning system is primarily based on that sanctioning system, it is better than having no trust at all. It is a different story, however, when in a social dilemma situation, without the presence of a sanctioning system, people already *do* trust each other to be internally motivated to cooperate. Then, the installation of a sanctioning system may undermine this existing trust and under these circumstances the group may be better off not having a sanctioning system at all.

So, a sanction's trust-undermining effect may well depend on the extent to which people already trust each other. If people distrust each other, a sanction can be an effective way to increase trust and then it might not matter that it concerns trust in others being externally motivated to cooperate. If people already trust each other without a coercive system (i.e., they have trust in others' internal motivation to cooperate), however, a sanction may be expected to undermine this trust. People differ in the extent to which they are inclined to trust other people (Sato & Yamagishi, 1986; Yamagishi, 1986, 1992): some people generally have high trust in others (high-trusters), whereas others generally have low trust in others (low-trusters). As such, we argue that a sanctioning system may increase trust in others being externally motivated to cooperate for some people, and at the same time undermine trust in others being internally motivated to cooperate for other people.

In our second study, we tested this differential effect of a sanctioning system on trust for high-trusters and low-trusters in the RTS paradigm. We expected a sanctioning system to undermine trust in others being internally

motivated to cooperate to a greater extent amongst high-trusters than amongst low-trusters (see for a similar reasoning with respect to social motives, De Dreu et al., 1998). Accordingly, we anticipated that after having removed the sanction in the sanction condition, high-trusters will show less trust in the sanction condition than in the no-sanction condition, whereas trust among low-trusters will be equally low in the sanction condition as in the no-sanction condition. We expected the same pattern for cooperation.

In addition, we wished to explore the factors underlying the effect of sanctions on trust more thoroughly. Although a sanction undermined trust in our first study, our data did not allow us to assess whether the sanctioning system created trust in others being externally motivated to cooperate (i.e., trust based on the sanctioning system), whether the sanctioning system caused people to attribute other people's behavior to the presence of the sanctioning system, and whether it gave people the idea that others mainly pursued their self-interest. Because these factors were expected to underlie the sanctioning system's undermining of trust that others are internally motivated to cooperate, we wished to obtain additional measures of these factors in Study 2. So, we assessed the extent to which people attributed behavior to the presence of the sanctioning system ("sanction attribution"), the extent to which they based their trust upon the sanctioning system ("trust that others are externally motivated to cooperate"), and the extent to which they thought people pursued their self-interest ("perceived motive of self-interest").

The perceived motive of self-interest was measured in both the sanction conditions and the no-sanction conditions. We expected it to be higher in the sanction conditions than in the no-sanction conditions. Contrary to perceived motive of self-interest, however, it is difficult to attempt to measure sanction attribution and trust in others being externally motivated to cooperate when there is no sanctioning system present at all. One way of dealing with this is by comparing sanctioning systems that differ in the extent to which they sanction defection. After all, a large sanction on defection punishes defection to a greater extent than a small sanction on defection. As a method of testing the effect of sanction on sanction attribution and trust in others being externally motivated to cooperate, we therefore created, in addition to the no-sanction condition, two different sanction conditions: a small-sanction condition and a large-sanction condition. Our expectations were that, if people based their trust on the sanctioning system and attribute behavior to the system, they would do so to a greater extent when the sanction is large than when the sanction is small. Therefore, we hypothesized that both trust in others being externally motivated to cooperate and sanction attribution would be higher in the large-sanction condition than in the small sanction condition.

Method

Design and participants

For this experiment with a 3 (sanction: no, small or large) \times 2 (dispositional trust: low vs. high) between-subjects design, 126 students of Leiden University participated. The experiment lasted 45 min and participants were paid EUR 5. Participants were divided randomly over the six conditions. Data from three participants were discarded because they had expressed doubts about the reality of the experimental social dilemma situation. This resulted in 123 participants (45 male and 78 female, mean age of 20.9 years).

Procedure

To distinguish high-trusters from low-trusters, participants answered the eight items of a trust questionnaire (Sato & Yamagishi, 1986; Yamagishi, 1986) on a scale ranging from 1 (*totally disagree*) to 7 (*totally agree*). An example of one item is “You should not trust other people unless you know them very well.” Higher scores indicated lower levels of trust. The trust questionnaire was sufficiently reliable ($\alpha = .75$) and a median split (*Median* = 3.50) was performed to create a group of high-trusters ($n = 65$) and a group of low-trusters ($n = 58$).

Next, similar to Study 1, two social dilemma situations (A and B) were presented to participants. This time, however, participants were told that each group participating in the experiment would be paid. Whether they were paid according to situation A or B, would be determined randomly after the experiment. Again, each chip was worth EUR 0.05. Now, the sanction was either EUR 0.50 (in the small-sanction conditions) or EUR 5 (in the large-sanction conditions). The rest of the procedure, the measure of overall trust in group members, and the measure of cooperation, were similar to Study 1, except for the added measures of trust that others are externally motivated to cooperate, sanction attribution, and perceived motive of self-interest. These measures were several statements which participants were asked to rate on a scale ranging from 1 (*totally disagree*) to 7 (*totally agree*). Statements regarding trust in others being externally motivated to cooperate and sanction attribution were measured after participants had made their decision in situation A in all sanction conditions. Trust in others being externally motivated to cooperate was measured by the statement “I trust that the sanctioning system urges other group members to contribute chips to the group.” Sanction attribution was measured by five items measuring the degree to which group members thought both their own and other group members’ actions were based on the presence of the sanctioning system (example: “I think the fine had an important influence on other group members’ choice”). These five items formed a reliable scale ($\alpha = .91$). “Perceived motive of self-interest” was measured after participants had

made their decision in situation B (after removal of the sanctioning system in the sanction conditions) by asking to what extent they agreed with the statement that other group members had based their decision mainly on self-interest (on a scale ranging from 1 [*absolutely not*] to 7 [*absolutely*]).

Finally, all participants were debriefed, thanked and paid EUR 5 for their participation. All participants agreed to this procedure.

Results

For reasons of clarity and to present the results in the same way over the experiments, we report the 3 (sanction) \times 2 (dispositional trust) \times 2 (phase) ANOVA using the distinction between low- and high-trusters that was created by median splitting the trust scale. We would like to note, however, that for all dependent variables in Study 2, we also performed regression analyses using the continuous trust scale. Compared to the corresponding ANOVA’s the regression analyses showed similar results.

Trust

To assess whether trust had decreased in phase 2 compared with phase 1 we performed a 3 (sanction) \times 2 (dispositional trust) \times 2 (phase) ANOVA on trust with phase as a within-subjects factor. This analysis yielded a main effect of phase ($F[1, 117] = 26.01, p < .001$), qualified by a Phase \times Sanction interaction ($F[2, 117] = 9.14, p < .001$). Trust in phase 2 was lower ($M = 4.52, SD = 1.63$) than trust in phase 1 ($M = 5.22, SD = 1.28$). The Phase \times Sanction interaction (see upper part of Table 2) shows that there is a significant decrease of trust in the large-sanction condition ($t[41] = 5.14, p < .001$), in the small-sanction condition ($t[40] = 2.50, p = .02$), but not in the no-sanction condition ($t[39] = 0.64, ns$).

To test whether in phase 2 trust in the sanction conditions had decreased below the level of the no-sanction conditions and to test whether this differed between low-trusters and high-trusters, a 3 (sanction) \times 2 (dispositional trust) ANOVA on trust in phase 2 was performed.

Table 2
Trust and cooperation as functions of phase and sanction, Study 2

	Mean		SD	
	Phase 1	Phase 2	Phase 1	Phase 2
<i>Trust</i>				
No sanction	5.15 ^a	5.05 ^a	1.48	1.45
Small sanction	5.02 ^a	4.56 ^{ab}	1.19	1.53
Large sanction	5.48 ^a	3.98 ^b	1.13	1.75
<i>Cooperation</i>				
No sanction	53.85 ^a	58.78 ^a	32.76	34.69
Small sanction	67.73 ^b	56.95 ^a	31.95	36.89
Large sanction	73.98 ^b	48.33 ^a	24.85	34.47

Note. For each row, different superscripts differ significantly ($p < .05$, paired-samples *t* test).

Table 3

Trust in fellow group members and cooperation in phase 2 as a function of sanction and dispositional trust, Study 2

Dispositional trust	Mean			SD		
	No sanction	Small sanction	Large sanction	No sanction	Small sanction	Large sanction
<i>Trust</i>						
Low	5.00 ^a	4.27 ^{ab}	4.20 ^{ab}	1.48	1.64	1.70
High	5.11 ^a	4.89 ^a	3.85 ^b	1.45	1.37	1.79
<i>Cooperation</i>						
Low	49.05 ^{ab}	58.23 ^{ab}	60.33 ^{ab}	34.81	37.77	34.20
High	69.53 ^a	55.47 ^{ab}	41.67 ^b	32.06	36.82	33.38

Note. Different superscripts differ significantly ($p < .05$, LSD).

This ANOVA demonstrated only a main effect of sanction, $F(2, 117) = 4.10$, $p = .02$. Post hoc tests indicated that trust was lower in the large-sanction condition ($M = 3.98$, $SD = 1.75$) than in the no-sanction condition ($M = 5.05$, $SD = 1.45$), LSD, $p < .05$. The level of trust in the small-sanction condition ($M = 4.56$, $SD = 1.53$) did not differ from the levels of trust in the no-sanction and large-sanction condition.

Cooperation

To investigate whether cooperation decreased in phase 2 compared with phase 1, as a function of sanction and dispositional trust, we performed a 3 (sanction) \times 2 (dispositional trust) \times 2 (phase) ANOVA on cooperation, with phase as a within-subjects factor. This analysis yielded a main effect of phase ($F[1, 117] = 16.45$, $p < .001$), qualified by a Phase \times Sanction interaction ($F[2, 117] = 11.89$, $p < .001$). There was less cooperation in phase 2 ($M = 54.60$, $SD = 35.37$) than in phase 1 ($M = 65.35$, $SD = 30.92$). The Phase \times Sanction interaction (see lower part of Table 2) shows that cooperation significantly decreased in the large-sanction condition ($t[41] = 4.85$, $p < .001$), in the small-sanction condition ($t[40] = 3.00$, $p = .005$), but not in the no-sanction condition ($t[39] = -1.29$, *ns*).

To test whether cooperation in the sanction condition had decreased below the level of the no-sanction condition and to test whether this differed between low-trusters and high-trusters, a 3 (sanction) \times 2 (dispositional trust) ANOVA on cooperation in phase 2 was performed. The ANOVA on cooperation demonstrated a significant Sanction \times Dispositional Trust interaction, $F(2, 117) = 3.13$, $p < .05$. This interaction shows that the presence of a sanctioning system in phase 1 decreased cooperation when dispositional trust was high and not when dispositional trust was low (see lower part of Table 3). Table 3 shows that for high-trusters, cooperation was lower in the large-sanction condition than in the no-sanction condition. The level of cooperation in the small-sanction condition fell in between and did not significantly differ from the no-sanction and large-sanction condition. For low-trusters, cooperation did not differ between the different sanction conditions. These results corroborate our reasoning that a sanction may undermine cooperation when initial trust is high.

Mediation analysis

So far, the data in Study 2 show that the sanction undermined trust in others being internally motivated to cooperate. Further, the data show that, among high-trusters, the sanction also decreased the level of cooperation in phase 2. Was this undermined cooperation the result of undermined trust? To investigate this possibility, we tested a possible mediation of phase 2 trust on the effect of sanction on phase 2 cooperation, and did this for low-trusters and high-trusters separately. We performed these analyses by estimating a series of regression models (Baron & Kenny, 1986).

For low-trusters a regression analysis on cooperation with sanction as predictor was performed. Sanction appeared to be a non-significant predictor for cooperation ($\beta = .13$, *ns*). So, for low-trusters, we could already conclude that there was no mediation because the sanction did not have an effect on phase 2 cooperation in the first place. Therefore, we did not need to perform any further steps to test mediation.

We turned to testing the mediation for high-trusters. First, a regression analysis on cooperation with sanction as predictor was performed. It showed that the former presence of the sanction predicted cooperation in phase 2, $\beta = -.33$, $p = .007$. Also, a regression analysis on phase 2 trust with sanction as predictor demonstrated that the former sanction predicted the mediator phase 2 trust, $\beta = -.33$, $p = .008$. In a regression analysis on cooperation with both phase 2 trust and sanction, phase 2 trust predicted cooperation ($\beta = .53$, $p < .001$), whereas the effect of sanction on cooperation disappeared and became non-significant ($\beta = -.16$, $p = .15$). Moreover, this decrease of the effect of sanction caused by the inclusion of the mediator phase 2 trust, was significant (Goodman I test value = -2.36 , $p = .02$). So, this provided support for the idea that the negative effect among high-trusters of the former presence of the sanction on phase 2 cooperation was mediated by phase 2 trust.²

² We also tested whether there was mediation for the complete design, so including both low-trusters and high-trusters. There was not (sanction did not significantly predict phase 2 cooperation, $\beta = .13$, *ns*).

Additional measures

We performed analyses on the additional measures of trust that others are externally motivated to cooperate, of sanction attribution and of the perceived motive of self-interest. Because trust in others being externally motivated to cooperate and sanction attribution could only be measured in the small- and large-sanction conditions, the no-sanction condition was left out of the analyses on these two measures.

Trust in others being externally motivated to cooperate.

As expected, a 2 (sanction: small vs. large) \times 2 (dispositional trust) ANOVA on trust in others being externally motivated to cooperate yielded only a main effect of sanction, $F(1, 79) = 8.37$, $p = .005$. This analysis showed that there was more trust in others being externally motivated to cooperate in the large-sanction condition ($M = 5.52$, $SD = 1.38$) than in the small-sanction condition ($M = 4.63$, $SD = 1.50$).

Sanction attribution. A 2 (sanction: small vs. large) \times 2 (dispositional trust) ANOVA on the sanction-attribution scale yielded a sanction main effect, $F(1, 79) = 23.07$, $p < .001$. Attributions to the sanctioning system were higher in the large-sanction condition ($M = 4.76$, $SD = 1.33$) than in the small-sanction condition ($M = 3.46$, $SD = 1.23$).

Perceived motive of self-interest. A 3 (sanction) \times 2 (dispositional trust) ANOVA on the perceived self-interest motive yielded only a significant main effect of sanction. As hypothesized, participants perceived their fellow group members to be motivated by self-interest motive to a greater extent in the large-sanction condition ($M = 5.14$, $SD = 1.39$) than in both the small-sanction condition ($M = 4.44$, $SD = 1.43$) and the no-sanction condition ($M = 4.28$, $SD = 1.75$), LSD post hoc tests, $p < .05$.

Discussion

The results of Study 2 again show that when an existing sanctioning system is removed, both trust and cooperation decrease. This was observed for both high-trusters and low-trusters. After removing the sanction, trust decreased below the level of trust in the no-sanction condition, supporting our hypothesis that trust that others are internally motivated to cooperate is undermined by the sanction. In contrast to our hypothesis, this did not particularly occur among high-trusters. However, in line with our hypothesis, the undermining of cooperation by a sanction did depend on dispositional trust: The sanction undermined cooperation only for high-trusters. A mediational analysis suggested that this was due to the undermined trust. Although we do not wish to argue that results of mediational analyses should be taken as

conclusive evidence, this finding does suggest that when initial trust is high, a sanction can undermine this trust and as a result undermine cooperation in a subsequent situation.

Moreover, in this study we found support for all of our supposed underlying processes. First, our findings show that a large-sanction brought about more trust in others being externally motivated to cooperate than a small sanction. This supports our assumption that a sanctioning system brings about trust that is based on the presence of that system. Second, a large sanction induced people to attribute fellow group members' behavior to a greater extent to the sanctioning system than a small sanction. This supports our argument that a sanction causes people to ascribe other people's behavior to the presence of the sanctioning system. Third, in the large-sanction condition perceived motives of self-interest were higher than in both the small-sanction condition and the no-sanction condition. This supports our expectation that a sanction induces people to think that other people mainly pursue their self-interest.³

Study 3

In Study 2, we showed that a sanctioning system undermines trust and undermines cooperation, particularly when the initial level of trust was high. For the assessment of this initial level of trust we used a measure of dispositional trust by distinguishing between low-trusters and high-trusters. One could argue, however, that these findings cannot instantly be generalized to other situations in which trust is initially high or low. After all, many factors contribute to the level of trust that people show. In addition to dispositional factors, trust can indeed be influenced by various social cues like, for example, group size, heterogeneity of the group, possibility for communication or possibility for social control (Dawes, 1980; Kopelman et al., 2002; Liebrand et al., 1992; Messick & Brewer, 1983; Messick et al., 1983). Therefore, we found it important not only to measure trust as a disposition, but also to manipulate people's initial level of trust.

³ It may be interesting to note that the small sanction, with regard to trust and cooperation, was always "in between" the no-sanction and the large-sanction condition: It seemed to undermine trust and cooperation to a lesser extent than the large sanction (and non-significantly). This suggests that the trust-undermining effect of a sanction is contingent upon the severity of the sanction. Because people may rely on a strong sanctioning system more than on a weak one, a strong sanctioning system may bring about extrinsic trust and replace internal trust to a greater extent than a weak sanctioning system. To exactly understand the role of the size of a sanction, more detailed research is necessary.

In Study 3, we manipulated trust by means of a newly developed trust manipulation. This manipulation was based on the findings that people adopt the behavior that others show (Schroeder, Jensen, Reed, Sullivan, & Schwab, 1983) and on the generally accepted notion that trust tends to evoke trust and distrust tends to evoke distrust (Blomqvist, 1997). To manipulate trust participants were, before they were presented the two social dilemma situations, exposed to a different mixed-motive situation in which they observed four persons who either trusted or distrusted another person. We anticipated our participants to adapt their level of trust to the level that was shown by these four persons. Again, in phase 2 after the sanction was removed, we expected lower trust and cooperation in the sanction condition than in the no-sanction condition, especially for the high-manipulated trust condition.

Method

Design and participants

For this experiment with a 2 (no sanction, sanction) \times 2 (manipulated trust: low vs. high) between-subjects design, 100 students (38 males and 62 females, mean age of 20.7) of Leiden University participated. The experiment lasted 45 min and participants were paid EUR 3.50. Participants were assigned randomly over the four conditions. Data from four participants were discarded because they had expressed doubts about the reality of the experimental social dilemma situation.

Procedure

The experiment consisted of two separate parts that were presented to participants as two unrelated experiments. First, trust was manipulated by presenting participants a trust game. Second, participants were presented the social dilemma choice situations similar to Studies 1 and 2.

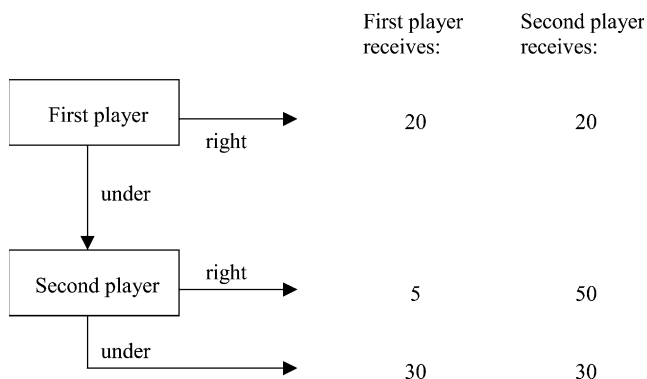


Fig. 1. Structure of trust game used as manipulation of trust in Study 3.

Manipulation of trust. To manipulate trust we presented participants with a situation known as the “Trust Game” (Dasgupta, 1988; Kreps, 1990; Malhotra & Murnighan, 2002). Participants were told that this trust game involved two players: player 1 and player 2. Player 1 had to decide whether to trust player 2. Player 2, if trusted, had to decide whether to honor or violate this trust. The exact situation is pictured in Fig. 1. To be more specific, it was first up to player 1 to decide whether to choose “right” or “under.” If player 1 chose right, both players would earn DFL 20 and the game would be finished. However, if player 1 chose under, it was then up to player 2 to decide whether to choose right or under. If player 2 chose right, he or she would earn DFL 50 whereas player 1 would earn DFL 5. If player 2 chose under, both players would earn DFL 30.⁴

Participants in our study (who we will refer to as “actual participants”) were told that in an experiment run the year before, participants in that experiment (who we will refer to as “former participants”) were presented with this situation and were asked to make a choice between “under” or “right.” The actual participants were shown the choices of four of the former participants who had the role of player 1 in the former experiment. They were told that these choices “reflected what most people did in these kind of situations.” Also, for each choice presented, they were shown a text typed by the former participants explaining the choice they made. In reality, both the four choices and the accompanying text were pre-programmed. Each text implied something about trusting (in the high-manipulated trust condition) or not trusting (in the low-manipulated trust condition) player 2.

In the low-manipulated trust condition actual participants were shown that each of the former participants had chosen “right.” In the high-manipulated trust condition actual participants were shown that each of the former participants had chosen “under.” An example of an accompanying text in the high-manipulated trust condition (the texts for the low-manipulated trust condition are given between brackets) was:

I want to earn money and the best way to do that is (not) to trust the other player and to choose under (right).

Then, the actual participants were asked four questions, measuring the extent to which they, if they were player 1, would trust player 2. These questions were answered on a scale ranging from 1 to 7, higher scores indicating higher levels of trust towards the second player. An example of these questions is “If I were player

⁴ At the time of the experiment, the euro (EUR) had just been introduced. Because in our trust manipulation, we showed some bogus behavioral choices of participants in an alleged experiment of the previous year (when the euro was not yet introduced), we described the pay-offs in Dutch guilders (DFL).

1, I would..." (1 = not trust player 2, 7 = trust player 2). These questions together formed the "trust game scale" ($\alpha = .92$).⁵

Social dilemma situations. Next, similar to Studies 1 and 2, two social dilemma situations (A and B) were presented to participants. The money earned in the experiment would be paid as an extra (in addition to the EUR 3.50) to one group and this group would be paid the money earned in either situation A or B. Now, in situation A, the group member who contributed the least, was sanctioned in the sanction conditions. This sanction meant that this group member would lose all his or her chips in that situation. The rest of the procedure was similar to Studies 1 and 2. Again, the additional measure of perceived motive of self-interest was measured.

Finally, all participants were debriefed, thanked and paid EUR 3.50 for their participation. One group was paid extra according to the choices made by its group members in situation B. All participants agreed to this procedure.

Results

Manipulation checks

A 2 (sanction) \times 2 (manipulated trust) ANOVA on the trust game scale showed only a main effect of trust ($F(1, 91) = 17.97, p < .001$), demonstrating a higher level of trust in the second player in the high-manipulated trust condition ($M = 4.67, SD = 1.66$) than in the low-manipulated trust condition ($M = 3.21, SD = 1.67$). We can thus conclude that the trust manipulation was successful.

Trust

To assess whether trust had decreased in phase 2, compared with phase 1, as a function of sanction and manipulated trust, we performed a 2 (sanction) \times 2 (manipulated trust) \times 2 (phase) ANOVA on trust with phase as a within-subjects factor. This yielded a main effect of phase ($F[1, 91] = 52.87, p < .001$), qualified by a Phase \times Sanction interaction ($F[1, 91] = 55.21, p < .001$). Trust in phase 2 was lower ($M = 4.25, SD = 1.72$) than trust in phase 1 ($M = 5.15, SD = 1.22$). The Phase \times Sanction interaction (see upper part of Table 4)

shows that there is a significant decrease of trust in the sanction condition ($t[48] = 7.80, p < .001$), but not in the no-sanction condition ($t[45] = 0.44, ns$).

We then tested whether, in phase 2, trust in the sanction condition had decreased below the level of the no-sanction condition and whether this differed between the high-manipulated trust and the low-manipulated trust condition by performing a 2 (sanction) \times 2 (manipulated trust) ANOVA on trust in phase 2. The ANOVA on phase 2 trust revealed main effects of manipulated trust ($F[1, 91] = 9.70, p = .002$) and, more important, sanction ($F[1, 91] = 34.70, p < .001$). The main effect of manipulated trust demonstrated that trust was lower in the low-manipulated trust condition ($M = 4.93, SD = 1.54$) than in the high-manipulated trust condition ($M = 3.77, SD = 1.77$). The main effect of sanction showed a lower level of trust in the sanction condition ($M = 3.41, SD = 1.66$) than in the no-sanction condition ($M = 5.15, SD = 1.28$). These results indicate that the sanction had undermined trust in others being internally motivated to cooperate.

Cooperation

We performed a 2 (sanction) \times 2 (manipulated trust) \times 2 (phase) ANOVA on cooperation with phase as a within-subjects factor to investigate whether cooperation too had decreased in phase 2 compared with phase 1 as a function of sanction and manipulated trust. This yielded a main effect of phase ($F[1, 91] = 22.22, p < .001$), qualified by a Phase \times Sanction interaction ($F[1, 91] = 26.72, p < .001$). There was less cooperation in phase 2 ($M = 46.35, SD = 34.77$) than in phase 1 ($M = 60.51, SD = 29.61$). The Phase \times Sanction interaction (see lower part of Table 4) shows that cooperation decreased significantly in the sanction condition ($t[48] = 6.11, p < .001$), and not in the no-sanction condition ($t[45] = -0.39, ns$).

To test whether in phase 2 cooperation in the sanction condition had decreased below the level of the no-sanction condition and whether this differed between the high-manipulated trust and the low-manipulated trust condition, we performed a 2 (sanction) \times 2 (manipulated trust) ANOVA on phase 2 cooperation. It showed a

Table 4
Trust and cooperation as functions of phase and sanction, Study 3

	Mean		SD	
	Phase 1	Phase 2	Phase 1	Phase 2
<i>Trust</i>				
No sanction	5.13 ^a	5.15 ^a	1.26	1.28
Sanction	5.16 ^a	3.41 ^b	1.20	1.66
<i>Cooperation</i>				
No sanction	52.65 ^a	53.91 ^a	32.05	31.12
Sanction	67.88 ^a	39.24 ^b	25.29	36.24

Note. For each row, different superscripts differ significantly ($p < .05$, paired-samples t test).

⁵ Some readers may wonder why we did not use the "investment game" as described by Berg, Dickhaut, and McCabe (1995). In the investment game of Berg et al. the truster can choose between a range of 0–10 units to invest, the investment is tripled and the trustee is asked how many units he wants to give to the truster. Because the relation between the trust shown by the truster and the level of trust perceived by the trustee did not seem to be entirely linear in this investment game, we were not entirely certain that showing choices of persons in the investment game would have unequivocal effects on perceived trust. Because in the dichotomous trust game there was no question that choosing "under" communicated trust and choosing "right" communicated distrust, we decided to use this game.

main effect of sanction, ($F[1, 91] = 4.22, p = .04$), qualified by a marginally significant Sanction \times Manipulated Trust interaction, $F(1, 91) = 2.78, p < .10$. The main effect of sanction indicated that cooperation was lower in the sanction condition ($M = 39.24, SD = 36.78$) than in the no-sanction condition ($M = 53.91, SD = 31.12$). To investigate the pattern of the cooperation data, we performed post hoc tests. These suggested that mainly in the high-manipulated trust condition cooperation was significantly higher with no experience of a sanctioning system in phase 1 than with the experience of a sanctioning system in phase 1. In the low-manipulated trust condition, the previous experience of a sanctioning had no effect: The cooperation level in the sanction condition did not significantly differ from that of the no-sanction condition (see Table 5). The results on phase 2 cooperation indicate that the sanction had undermined cooperation and that this was especially the case when trust was initially high rather than low.

Mediation analysis

The data in Study 3 show that the sanction undermined trust in others being internally motivated to cooperate. Further, the data also show that the sanction decreased the level of cooperation in phase 2. Although this was an overall effect, the pattern of data suggest that this was especially true when manipulated trust was high. We investigated the possible mediating role of phase 2 trust on the effect of sanction on cooperation for both when manipulated trust was low and manipulated trust was high and performed the corresponding analyses by estimating a series of regression models (Baron & Kenny, 1986).

For the low manipulated trust conditions a regression analysis on cooperation with sanction as predictor was performed. Sanction appeared not to predict cooperation ($\beta = -.04, ns$). So we could already conclude that there was no mediation in the low manipulated trust conditions because the sanction did not have an effect on phase 2 cooperation in the first place. Therefore, we did not need to perform any further steps to test mediation.

In the high manipulated trust conditions a regression analysis on cooperation with sanction as predictor showed that the former presence of the sanction predicted cooperation in phase 2, $\beta = -.37, p = .009$. Also, a regression analysis on phase 2 trust with sanction as predictor demonstrated that the former sanction predicted

the mediator phase 2 trust, $\beta = -.62, p < .001$. In a regression analysis on cooperation with both phase 2 trust and sanction as predictors, phase 2 trust predicted cooperation ($\beta = .66, p < .001$), whereas the effect of sanction on cooperation disappeared and became non-significant ($\beta = .03, p = .83$). Moreover, this decrease in the effect of sanction caused by the inclusion of the mediator phase 2 trust, was significant (Goodman I test value $-2.13, p = .03$). This gave support to the idea that, in the high-manipulated trust condition, the negative effect of the sanction on phase 2 cooperation was mediated by phase 2 trust.⁶

Perceived motive of self-interest

A 2 (sanction) \times 2 (manipulated trust) ANOVA on the additional measure of the perceived motive of self-interest yielded main effects of sanction ($F[1, 92] = 24.73, p < .001$) and manipulated trust ($F[1, 92] = 5.94, p = .02$). Like Study 2, participants appeared to perceive self-interest to be a motive of fellow group members to a greater extent in the sanction condition ($M = 5.22, SD = 1.52$) than in the no-sanction condition ($M = 3.63, SD = 1.62$), and to a greater extent in the low-manipulated trust condition ($M = 4.85, SD = 1.60$) than in the high-manipulated trust condition ($M = 4.06, SD = 1.83$).

Discussion

This study replicates the basic findings of Studies 1 and 2: When a sanctioning system is removed, trust decreases below the level of trust demonstrated by people who had no previous experience with a sanctioning system. This undermining of trust in others being internally motivated to cooperate occurred regardless of whether manipulated trust was low or high. Interestingly, the sanctioning system also undermined cooperation and seemed particularly true when initial trust was high. Again, the sanction caused people to perceive that their fellow group members were motivated by self-interest.

General discussion

All three studies show that a sanctioning system can undermine trust in others. Of course, the presence of the sanctioning system may have increased trust in others

Table 5

Cooperation in phase 2 as a function of sanction and manipulation trust, Study 3

Manipulation trust	Mean		SD	
	No sanction	Sanction	No sanction	Sanction
Low	42.37 ^a	39.68 ^a	28.12	39.52
High	64.50 ^b	38.79 ^a	30.48	36.53

Note. Different superscripts differ significantly ($p < .05, LSD$).

⁶ We also tested whether there was mediation using the complete design, so including both low and high manipulated trust conditions. There appeared to be a full mediation. There was a significant effect of sanction on phase 2 trust, $\beta = -.51, (p < .001)$ and when controlling for phase 2 trust the effect of sanction on phase 2 cooperation changed from $\beta = -.21 (p = .04)$ to $\beta = .16 (p = .09)$. This mediation appeared to be significant (Goodman I test value = $-4.61, p < .001$).

being *externally* motivated to cooperate. But at the same time, the sanctioning system made participants attribute their fellow group members' behavior to the sanctioning system and generated the idea that group members were guided by the motive of self-interest. Consequently, removing the sanction showed that the sanctioning system decreased trust below the level of trust of people who had not experienced the presence of a sanctioning system. All of this indicates that the sanction actually decreased the trust in fellow group members being internally motivated to cooperate.

We also hypothesized that both trust in others being internally motivated to cooperate and cooperation itself would be undermined especially when trust was initially high rather than low. Interestingly, in both Studies 2 and 3 the sanction undermined trust in others being internally motivated to cooperate *both* when initial trust was low *and* when initial trust was high. We can therefore conclude that a sanction is in general detrimental for the level of trust in others being internally motivated to cooperate and that even if the initial level of trust is low anyway, a sanctioning system can decrease trust even further.

The present research also addressed the question of whether these detrimental effects on trust also affect behavior. In Study 1 of this paper, in which we did not take the initial trust level into account, the sanction seemed not to undermine cooperation. There was a relation between trust and cooperation, but apparently the relation was not strong enough for the sanction having the same effect on cooperation as it had on trust. The data in Study 1 do not allow conclusive answers to why this was the case, but the data of Studies 2 and 3 can help to clarify: Studies 2 and 3 varied initial trust levels and show that the sanction *did* undermine cooperation but only when trust was initially high. So, when trust is initially at a high level (whether this is due to a personality trait or due to situational cues), having experienced a sanctioning system undermines cooperation. Although we realize that mediational analyses should not automatically be regarded as indisputable evidence for mediation, our mediation analyses did support the idea that the negative effect of a sanction on cooperation when initial trust was high was mediated by undermined trust. Thus, the sanction appears to have caused a decline in trust in other people being internally motivated to cooperate, which became visible after the sanction had been removed. In turn, this undermined trust seems to have caused a decline in cooperation when the initial level of trust was high. When trust was initially low, however, trust in others being internally motivated to cooperate also decreased as a result of the sanction, but this did not result in lower levels of cooperation. A similar process may have occurred in Study 1. In Study 1, we did not measure or manipulate initial trust, but maybe only participants with a high initial level of trust based their

behavior on their undermined trust, so that there was only a sanction effect on cooperation for those initially high in trust and not for those initially low in trust.

The next question then is, of course, how to account for these differential effects as a function of trust level. A possible explanation for the differential effect of the sanction among people who were initially high versus low in trust, might be the fact that the information provided by the sanction is more unexpected for people initially high in trust than for people initially low in trust. Research of Pillutla and Chen (1999) has shown that if participants in a social dilemma situation learned that others' behavior was consistent with expectations it did not affect participant's behavior, whereas if others behaved inconsistently with expectations it did. In our studies, the idea brought about by a sanctioning system that other people were not to be trusted might have been especially unexpected for participants with a high level of trust. The notion that others may be untrustworthy may have caused surprise, disappointment, anger (Olson & Janes, 2002), and vigilance (Olson, Roese, & Zanna, 1996). This negative arousal may have motivated people to think twice about their behavioral choice, leading to lower levels of cooperation. When trust was low to begin with, however, it is true that a sanction decreased trust even further, but because people were not trusting anyway this may have been in line with their expectations. Consequently, the impact might have been lower and may have affected cooperation to a lesser extent than when the initial level of trust was high. It may be worthwhile addressing the possible effects of the initial trust level on the behavioral consequences of trust information in future research.

The adverse effect of trust increasing measures

Some earlier literature already pointed out that certain measures imposed to increase trust may unintentionally harm trust. In the organization literature, Sitkin and Roth (1993) and Sitkin and Bies (1993) suggested that legalistic remedies to increase trust can have adverse effects (Kramer, 1999). Malhotra and Murnighan (2002) have shown that contracts can have a detrimental effect on interpersonal trust. The studies presented in this paper contribute to the knowledge about the negative effect of such measures on trust in various ways. First, we have empirically tested and confirmed that a measure to increase trust (i.e., a sanctioning system) can decrease trust that others are internally motivated to cooperate. Second, whereas Malhotra and Murnighan showed that a contract proposed by a trustee could undermine trust of the truster, we showed that a sanctioning system that is installed by an independent super ordinate entity also undermines trust in trustees. This suggests that it is the sanctioning system itself, instead of the act of proposing a contract (Malhotra & Murnighan, 2002) or the act of

punishing (De Dreu et al., 1998; Fehr & List, 2002) that is responsible for decreasing trust. Apparently, the mere presence of a sanctioning system can give people the idea that others mainly act in their own self-interest and are therefore not to be trusted. The third contribution of the studies presented in this paper is that they demonstrate that the detrimental effects of a sanction are not limited to trust, but also concern cooperative behavior in a social dilemma decision.

Dependency on a sanctioning system

The research presented in this paper, advocates that the mere presence of a sanctioning system can make people dependent on that sanctioning system: because trust that others will cooperate without the pressure of a sanctioning system is undermined, a sanctioning system needs to be kept in place to ensure cooperation.

A simulation study of Macy (1993) suggests otherwise. Macy's simulation model showed a pattern of increased contribution to a sanctioning system when cooperation rates were low. This resulted in higher cooperation rates. When these high cooperation rates had stabilized, contribution rates to a sanctioning system disappeared. Thus, this model suggests that, having brought about enough cooperation, people will remove a sanctioning system without a decline in cooperation rates. However, we have two reasons to think that this conclusion is premature. First, Macy's model did not include attributions of cooperation and the possible effect of the sanctioning system on trust in others. If it had, his model might predict a decreased level of trust that others would cooperate in absence of a sanctioning system and, as a result, contributions to the sanctioning system may have sustained. Second, we argue that in Macy's model the sanctioning system was never actually "removed." Although contribution to the system decreased when a high level of cooperation was attained, there was still opportunity to contribute to the sanctioning system. Thus, the model depicts a situation in which the opportunity to sanction is used less after having attained high levels of cooperation, rather than a situation in which the opportunity to sanction is entirely eliminated. In fact, recent research has shown that once a sanctioning system is installed, people do not want to get rid of it (Mulder, Van Dijk, De Cremer, & Wilke, in press).

Another remark with regard to the dependency of a sanctioning system, is that, in our experimental set-up, the sanction was only kept in place for one trial. With regard to trust in others, one could argue that a sanction needs to be left in place for a while to be effective: people may need time to get used to the sanction or they may need experience with it to see that others *can* be trusted. Would keeping the sanction in place for a longer time have undermined trust in our experiments as well? Possi-

bly, but note that one could also argue for the opposite effect. By keeping a sanctioning system for a long time people may learn to rely on the sanctioning system rather than rely on each other's good intentions. Seeing other people cooperate may not help as people attribute their cooperation to the presence of the sanctioning system (as they did in Study 2). Consequently, time may increase dependence on that system. Over time people may get used to the presence of a sanctioning and start to regard it as something normal and indispensable.

Support for the trust-undermining effect of a long-term sanctioning system, is reported by Yamagishi (1988a, 1988b). In his experiments he found that in a social dilemma without a sanctioning system, Japanese participants are less cooperative than American participants. Yamagishi argued that this was due to the fact that Japanese society was characterized to a greater extent by surveillance and monitoring systems than the American society. This suggests that being used to the presence of a sanctioning system decreases trust in others in a situation in which a sanctioning system is absent.

Of course, more research is needed to investigate to what extent a sanctioning system makes people depend on that system and the exact influence in this of the duration of a sanctioning system.

The detrimental effects of a sanction

Most earlier research on sanctioning systems in social dilemmas has shown direct positive effects on cooperation (Caldwell, 1976; Eek et al., 2002; Fehr & Gächter, 2002; McCusker & Carnevale, 1995; Tenbrunsel & Messick, 1999; Van Vugt & De Cremer, 1999; Wit & Wilke, 1990; Yamagishi, 1986, 1992) whereas in this paper we have focused on the negative effects of sanctioning systems. As we mentioned in our introduction, Tenbrunsel and Messick (1999) also showed that a sanctioning system can have negative effects, but then on the perceived decisional frame used in a social dilemma. Their study showed that sanctioning systems, whether weak (small sanction with a low chance of being caught) or strong (large sanction with a high chance of being caught) can negatively affect people's decisional frame, but that this only results in non-cooperative behavior when the sanction is weak and not when it is strong.

Whereas Tenbrunsel and Messick (1999) focused on a sanction's detrimental effect on people's own motives, we focused on its detrimental effects on trust in others. Moreover, whereas the large sanction in their study did not form a problem for the cooperation level, it did so in our studies. In Studies 2 and 3, the large sanction had harmful effects in a way that it caused defection when it was no longer present (when trust initially had been high). Thus, in this paper we have shown that also a large sanction can have detrimental effects on cooperation. In our paradigm, the behavioral effect came to light

when the sanctioning system was removed. However, there may be other ways in which adverse behavioral effects can manifest themselves. For example, when a sanctioning system is present, but behavior is insufficiently monitored, distrusting group members might be motivated to defect without being sanctioned, particularly when the opportunity comes up. Also, people might seek alternative ways of defection to avoid the sanction (Mulder, Van Dijk, De Cremer, & Wilke, 2005).

Another way in which the negative sanction effect can manifest itself is that undermined trust may generalize to behaviors other than those on which the sanctioning system focuses. For example, a sanction on exceeding the herring quota in the fishing industry, may not only undermine trust that other fishermen are internally motivated to restrain themselves from exceeding the herring quota but also undermine trust that other fishermen are internally motivated to restrain themselves from exceeding the cod quota. Also, it may undermine trust that fishermen in different areas without a sanctioning system on exceeding fish quota are internally motivated to restrain themselves from exceeding these fish quota. Further research using the RTS paradigm could shed some light on the question whether a sanction also undermines trust in different social dilemma situations than it concerns.

Concluding remarks

In this paper, we used the removal of the sanction as a means to show that the sanction undermined trust that others are internally motivated to cooperate. Our findings suggest that this adverse effect of a sanctioning system may not show itself directly in people's cooperation level as long as the sanctioning system exists, but may surface in unexpected ways and as such form a problem indeed. It is our hope that future research can further improve the understanding the paradoxical relation between sanctioning systems, trust, and cooperation.

References

- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.
- Berg, J., Dickhaut, J. W., & McCabe, K. A. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*, 122–142.
- Blomqvist, K. (1997). The many faces of trust. *Scandinavian Journal of Management, 13*, 271–286.
- Bruins, J. J., Liebrand, W. B., & Wilke, H. A. (1989). About the saliency of fear and greed in social dilemmas. *European Journal of Social Psychology, 19*, 155–161.
- Caldwell, M. D. (1976). Communication and sex effects in a five-person prisoner's dilemma game. *Journal of Personality and Social Psychology, 33*, 273–280.
- Cialdini, R. B. (1996). Social influence and the triple tumor structure of organizational dishonesty. In D. M. Messick & A. E. Tenbrunsel (Eds.), *Codes of conduct: Behavioral research into business ethics*. New York: Russel Sage Foundation.
- Coombs, C. H. (1973). A reparameterization of the prisoner's dilemma game. *Behavioral Science, 18*, 424–428.
- Dasgupta, P. (1988). Trust as commodity. In D. Gambetta (Ed.), *Trust, making and breaking cooperative relations* (pp. 49–72). Oxford: Basil Blackwell.
- Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology, 31*, 169–193.
- De Cremer, D., Dewitte, S., & Snyder, M. (2001). 'The less I trust, the less I contribute (or not)?' The effects of trust, accountability and self-monitoring in social dilemmas. *European Journal of Social Psychology, 31*, 93–107.
- De Cremer, D., & Van Dijk, E. (2002). Reactions to group success and failure as a function of identification level: A test of the goal-transformation hypothesis in social dilemmas. *Journal of Experimental Social Psychology, 38*, 435–442.
- De Dreu, C. K. W., Giebels, E., & Van de Vliert, E. (1998). Social motives and trust in integrative negotiation: The disruptive effects of punitive capability. *Journal of Applied Psychology, 83*, 408–422.
- Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology, 18*, 105–115.
- Deci, E. L., Benware, C., & Landy, D. (1974). The attribution of motivation as a function of output and rewards. *Journal of Personality, 42*, 652–667.
- Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*, 627–668.
- Eek, D., Loukopoulos, P., Fujii, S., & Gärling, T. (2002). Spill-over effects of intermittent costs for defection in social dilemmas. *European Journal of Social Psychology, 32*, 801–813.
- Fehr, E., & Falk, A. (2001). Psychological foundations of incentives. *European Economic Review, 46*, 687–724.
- Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*, 137–140.
- Fehr, E., & List, J. A. (2002). The hidden costs and returns of incentives: Trust and trustworthiness among CEOs. Unpublished manuscript.
- Frey, B. S. (1993). Motivation as a limit to pricing. *Journal of Economic Psychology, 14*, 635–664.
- Frey, B. S. (2000). Morality and rationality in environmental policy. *Journal of Consumer Policy, 22*, 395–417.
- Frey, B. S., & Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *The American Economic Review, 87*, 746–755.
- Gneezy, U., & Rustichini, A. (2000). A fine is a price. *Journal of Legal Studies, 29*, 1–17.
- Hobbes, T. (1651/1909). *Leviathan*. Oxford: Clarendon.
- Kerr, N. L. (1983). Motivation losses in small groups: A social dilemma analysis. *Journal of Personality and Social Psychology, 45*, 819–828.
- Kopelman, S., Weber, J. M., & Messick, D. M. (2002). Factors influencing cooperation in commons dilemmas: A review of experimental psychological research. In E. Ostrom, T. Dietz, P. C. Dolsak, P. C. Stern, S. Stonich, & E. U. Weber (Eds.), *The drama of the commons* (pp. 113–156). Washington, DC: National Academy Press.
- Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology, 50*, 569–598.
- Kreps, D. M. (1990). Corporate culture and economic theory. In J. E. Alt & K. A. Shepsle (Eds.), *Perspectives on positive political economy* (pp. 90–143). Cambridge: Cambridge University Press.
- Liebrand, W. B. G., Messick, D. M., & Wilke, H. A. M. (Eds.). (1992). *Social dilemmas: Theoretical issues and research findings*. Oxford: Pergamon Press.

- Luhmann, N. (1988). Familiarity, confidence, trust: Problems and alternatives. In D. Gambetta (Ed.), *Trust making and breaking relationships*. Oxford: Basil Blackwell.
- Macy, M. W. (1993). Backward-looking social control. *American Sociological Review*, 58, 819–836.
- Malhotra, D., & Murnighan, J. K. (2002). The effects of contracts on interpersonal trust. *Administrative Science Quarterly*, 47, 534–559.
- McCusker, C., & Carnevale, P. J. (1995). Framing in resource dilemmas: Loss aversion and the moderating effects of sanctions. *Organizational Behavior and Human Decision Processes*, 61, 190–201.
- Messick, D. M., & Brewer, M. B. (1983). Solving social dilemmas: A review. In L. Wheeler & P. Shaver (Eds.), *Review of personality and social psychology* (pp. 11–44). Beverly Hills, CA: Sage.
- Messick, D. M., Wilke, H. A. M., Brewer, M., Kramer, R. M., English Zemke, P., & Lui, L. (1983). Individual adaptations and structural change as solutions to social dilemmas. *Journal of Personality and Social Psychology*, 44, 294–309.
- Mulder, L. B., Van Dijk, E., De Cremer, D., & Wilke, H. A. M. (2005). When sanctions fail to increase cooperation in social dilemmas: Considering the presence of an alternative defection option. Manuscript submitted for publication.
- Mulder, L. B., Van Dijk, E., De Cremer, D., & Wilke, H. A. M. (in press). The effect of feedback on support for a sanctioning system in a social dilemma: A question of installing or maintaining the sanction. *Journal of Economic Psychology*.
- Olson, J. M., & Janes, L. M. (2002). Asymmetrical impact: Vigilance for differences and self-relevant stimuli. *European Journal of Social Psychology*, 32, 383–393.
- Olson, J. M., Roese, N. J., & Zanna, M. P. (1996). Expectancies. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 211–238).
- Parks, C. D., Henager, R. F., & Scamhorn, S. D. (1996). Trust and reactions to messages of intent in social dilemmas. *Journal of Conflict Resolution*, 40, 134–151.
- Pillutla, M. M., & Chen, X. P. (1999). Social norms and cooperation in social dilemmas: The effects of context and feedback. *Organizational Behavior and Human Decision Processes*, 78, 81–103.
- Pruitt, D. G., & Kimmel, M. J. (1977). Twenty years of experimental gaming: Critique, synthesis, and suggestions for the future. *Annual Review of Psychology*, 28, 363–392.
- Rapoport, A., & Eshed Levy, D. (1989). Provision of step-level public goods: Effects of greed and fear of being gyped. *Organizational Behavior and Human Decision Processes*, 44, 325–344.
- Robbins, T. L. (1995). Social loafing on cognitive tasks: An examination of the “sucker effect”. *Journal of Business and Psychology*, 9, 337–342.
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55, 68–78.
- Sato, K., & Yamagishi, T. (1986). Psychological factors in the public good problem: Free-riding and the lack of trust. *Japanese Journal of Experimental Social Psychology*, 26, 89–95.
- Schnake, M. E. (1991). Equity in effort: The “sucker effect” in co-acting groups. *Journal of Management*, 17, 41–55.
- Schroeder, D. A., Jensen, T. D., Reed, A. J., Sullivan, D. K., & Schwab, M. (1983). The actions of others as determinants of behavior in social trap situations. *Journal of Experimental Social Psychology*, 19, 522–539.
- Sitkin, S. B., & Bies, R. J. (1993). The legalistic organization: Definitions, dimensions, and dilemmas. *Organization Science*, 4, 345–351.
- Sitkin, S. B., & Roth, N. L. (1993). Explaining the limited effectiveness of legalistic “remedies” for trust/distrust. *Organization Science*, 4, 367–392.
- Strickland, L. H. (1958). Surveillance and trust. *Journal of Personality*, 26, 200–215.
- Tenbrunsel, A. E. (1999). Trust as an obstacle in environmental-economic disputes. *American Behavioral Scientist*, 42, 1350–1367.
- Tenbrunsel, A. E., & Messick, D. M. (1999). Sanctioning systems, decision frames, and cooperation. *Administrative Science Quarterly*, 44, 684–707.
- Van Vugt, M., & De Cremer, D. (1999). Leadership in social dilemmas: The effects of group identification on collective actions to provide public goods. *Journal of Personality and Social Psychology*, 76, 587–599.
- Wit, A., & Wilke, H. A. (1990). The presentation of rewards and punishments in a simulated social dilemma. *Social Behaviour*, 5, 231–245.
- Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, 51, 110–116.
- Yamagishi, T. (1988a). Exit from the group as an individualistic solution to the free rider problem in the United States and Japan. *Journal of Experimental Social Psychology*, 24, 530–542.
- Yamagishi, T. (1988b). The provision of a sanctioning system in the United States and Japan. *Social Psychology Quarterly*, 51, 265–271.
- Yamagishi, T. (1992). Group size and the provision of a sanctioning system in a social dilemma. In W. B. G. Liebrand, D. M. Messick, & H. Wilke (Eds.), *Social dilemmas: Theoretical issues and research findings. International series in experimental social psychology* (pp. 267–287).