**From spontaneous cooperation to spontaneous punishment –**

**Distinguishing the underlying motives driving spontaneous behavior in first and second order public goods**

Dorothee Mischkowski1,2, Andreas Glöckner2,3, Peter Lewisch4

1 University of Göttingen

2 University of Hagen

3 Max Planck Institute for Research on Collective Goods

4 University of Vienna

Total word count: 6491

**Abstract**

The spontaneous cooperation effect suggests a prosocial default in social dilemmas. We test whether this effect can be generalized to second order public goods, i.e., punishment behavior, and investigate the underlying motivations. In two studies, involving an iterative and a one-shot public goods game with a second-party punishment option, we measured Social Value Orientation before capturing response time of cooperation and punishment decisions as well the affective reactions upon. In line with spontaneous cooperation, we find that the invested resources to punish decrease over time. Negative affect moderates spontaneous punishment in that punishment decisions are quicker for persons upset about the contribution behavior of their group members. Unlike spontaneous cooperation, spontaneous punishment is not driven by dispositional prosocials, but by situational above-average cooperation. Most importantly, we find in an overall-analysis of our data a significant three-way interaction in that spontaneous punishment is valid for above-average, highly upset contributors.

*Keywords:* Spontaneous cooperation; spontaneous punishment, second-party punishment; Social Value Orientation; above-average contribution, negative affect**Introduction**

The question whether humans are intuitively good or evil is part of the discussion in many disciplines, i.e., philosophy, economics, psychology and beyond. Research on situations, where the individual gain stands in conflict with the communality’s welfare (e.g., in social dilemmas) reveal insights in the psychological mechanisms of cooperation behavior and social preferences (e.g., Hardin, 1968). Especially, the addition of process tracing methods in comparison to a pure choice analysis has contributed how and when people make decisions more or less in favor of the group or their own benefit (De Dreu et al., 2010; Fiedler, Glöckner, Nicklisch, & Dickert, 2013; Rand, Greene, & Nowak, 2012; Rand et al., 2014). As a major contribution to the field, Rand et al. (2012) suggested that cooperation behavior is not only conducted spontaneously (i.e., being related to shorter decision times), but can also be enhanced through intuitive processing manipulations (e.g., time pressure). In this regard, the answer clearly indicated good news that humans are indeed “of a good kind” and that prosociality can be even further promoted. However, the evidence is currently highly debated (e.g., as being part of a replication project by the Open Science Framework <https://osf.io/scu2f/>). So far, several replications and related projects have found evidence for (Nielsen, Tyran, & Wengström, 2014; Rand et al., 2014) and against the spontaneous cooperation effect (Tinghög et al., 2013; Verkoeijen & Bouwmeester, 2014). Additionally, several moderators of the effect have been identified, mainly on the field of personality characteristics, i.e., Social Value Orientation (Mischkowski & Glöckner, 2016), Honesty-Humility (Kieslich & Hilbig, 2014; Mischkowski & Glöckner, 2016) and trust in the cooperativeness of daily life interaction partners (Rand et al., 2012). They indicate spontaneous cooperation as the default for persons with a prosocial, honest or trusting attitude, but inapplicable for their counterparts.

This paper approaches the spontaneous cooperation debate from a different perspective, questioning its generalizability for punishment decisions, also known as second order public goods. Most importantly, it investigates whether punishment implies the same mechanisms as cooperation behavior, resulting in the same behavioral pattern as spontaneous cooperation. If so, the driving forces are less clear, as spontaneous punishment could consist of a revenge-oriented, affect-driven behavior as well as it could be trait-dependent – in line with spontaneous cooperation for prosocials – on SVO to restore justice.

*Punishment behavior in social dilemmas*

Fehr and Gächter (2002) show that third-party punishment (also called altruistic punishment, AP) is suited and used to enhance cooperation in repeated games. In punishment decisions, participants invest their own money to reduce the outcome of other players. In third-party settings, punishers are not part of the interacting group, e.g., the PGG, whereas in second-party settings, they are. Generally, AP represents a second order public good as ‚everybody in the group will be better off if free riding is deterred, but nobody has an incentive to punish the free riders’ (Fehr & Gächter, 2002, p. 137).

As underlying motivations driving punishment behavior, Fehr and Gächter (2002) point out the importance of emotions, being “an important proximate factor behind altruistic punishment“ (p. 139). In one-shot settings, they mainly capture the retributive character of punishing free-riders; in repeated settings this confounds with a future-oriented motivation to deter from further free-riding.[[1]](#footnote-1) The emotional arousal driving punishment behavior is likely to be related with differences in contribution behavior (e.g., the anger of having been exploited). Above-average contributors are identified as persons mainly engaging in altruistic punishment (Fehr & Fischbacher, 2004; Fehr, Fischbacher, & Gächter, 2002; Fehr & Gächter, 2002). From a trait perspective, cooperation behavior is linked with social preferences, i.e., SVO, being the stable concern for fairness and equality in outcomes (e.g., Van Lange, 2000). SVO has been found to be an important predictor of cooperation behavior, capturing a meta-analytical correlation of *r* = .30 (Balliet, Parks, & Joireman, 2009), or, once correcting for publication bias, an estimated *r* = .25 (Renkewitz, Fuchs, & Fiedler, 2011). So far, the correlation of SVO and punishment behavior has found less attention, although Bogaert, Boone, and Declerck (2008) explicitly refer to the emotions that are activated in prosocials, e.g., “feeling disgust when fairness has been violated (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003) [and] feeling pleasure when one engages in altruistic punishment (De Quervain, Fischbacher, Treyer, & Schellhammer, 2004) (…).” Bogaert et al. (2008) further point out a need for future research as they “[…] expect that such affective feelings might play an important role in guiding their behavior.” (p. 472). We follow their call and test the relation between prosociality, affect and punishment behavior over the course of response time.

To summarize, the aim of the current paper is twofold, first to analyze the course of punishment behavior over time that could be in line with the spontaneous cooperation effect for different reasons: Assuming a dominating prosocial motivation to punish, i.e., to restore equality in outcomes, we would expect an interaction with SVO. On the other hand, also retributive, affect-driven punishment decisions are likely to be related with shorter decision times. An interaction effect with (negative) affect would therefore underline the retributive character of spontaneous punishment. We therefore postulate the following hypotheses:

H1: Overall, punishment decisions are conducted quicker than non-punishment decisions (main effect)

H2: Analogue to spontaneous cooperation, spontaneous punishment is driven by prosocials. For proselfs, no difference in punishment behavior is expected over time (interaction effect)

H3: Spontaneous punishment is driven by negative affect (interaction effect).

**Study 1**

**Method**

*Materials.* Note that all materials and datasets are available online at https://osf.io/9rpwn/. In a longitudinal design, we measured SVO up to 12 hours before the lab session, taking the extended 15 items version of the SVO Slider Measure (Murphy, Ackermann, & Handgraaf, 2011). Importantly, the SVO Slider Measure allows for a distinction of prosociality in Joint Gain Maximization and Inequality Aversion (Ackermann & Murphy, 2012) that was captured to further distinguish the broad shield of prosocials. Particularly, the aspiration for joint gain maximization is likely to be incompatible with punishment behavior as resources on both sides are destroyed. For inequality averse persons, however, it is perfectly rational to punish as it enables them to reduce inequality.[[2]](#footnote-2)

Two measure cooperation behavior we conducted an iterative (two rounds) PGG in groups of four players (MPCR = .5), taking the instructions of Mischkowski and Glöckner (2016) that result from a forward-backwards translation of the instructions used by Rand et al. (2012, Study 1). As important extension, we added a punishment option after each round with a cost-to-impact ratio of 1:4. In order to allow for direct deterrence and reputation building, no stranger matching was established after the first round, resulting in the same group composition. We measured decision time of contribution decisions and – most importantly – of punishment decisions.

In order to capture the affective states, we relied on the PANAS (Krohne, Egloff, Kohlmann, & Tausch, 1996), explicitly instructed to be related to the contribution behavior of the group members and eight further, more specific affect items that captured the affect related to punishment decision (see instructions).[[3]](#footnote-3) In order not to interfere in the game process but to get a clean investigation of cooperation and subsequent punishment behavior, affect was captured after the PGG. Specifically, the reasons not to inquire affect between contributions and punishment decisions most importantly refer to avoid ‚affective labeling‘ (Lieberman et al., 2007) and ‚affective asynchrony‘ (Peters & Slovic, 2007); the former indicating that reporting one’s affect might reduce the affective state per se whereas affective asynchrony refers to the phenomenon that deliberation about one’s affective state might reduce the relation its spontaneously associated behavior.

*Procedure.* For an overview of the procedure in both studies, see Figure 1. Participants entered the lab having already completed the SVO Slider Measure and a demographic questionnaire concerning age, gender and education. In the lab, they first provided informed consent before getting hard copies of the PGG instructions and answering several control questions to assure game comprehension (see Rand et al., 2012).[[4]](#footnote-4) Note that instructions entailed information of a subsequent punishment option, so participants were aware of that opportunity for all group members. The PGG was conducted in BoXS (Bonn eXperimental System; Seithe, Morina, & Glöckner, 2015). Participants were matched in groups of four, being always in an anonymous setting as two to three groups were working in parallel. Participants made their contribution decisions first, before estimating the average contributions of their group members (beliefs). Afterwards, the display of feedback of the group members’ contributions was provided (see Figure 1). The second-party punishment option was divided in two slides. First, a binary punishment decision was requested („Do you want to punish at least one of your team members?“) that was relevant for decision time measurement. In case of a positive answer, the specification of the amount of punishment followed on a subsequent screen.

As indicated, at this point the measurement of affective states followed. Afterwards, a free-form field to justify one’s (non-) punishment decision was inserted in order to be able to classify the underlying motivations of punishment also on a qualitative level. In line with Rand et al. (2012) trust in the cooperativeness of daily life interaction partners and the previous experience with the experimental setting of social dilemma games (referred to as money splitting tasks) were captured at the end, before participants received feedback about their final outcome, were paid and debriefed.

*Participants.* An a priori power analysis using G\*Power (Faul, Erdfelder, Buchner, & Lang, 2009) revealed a required sample size of *N* = 114 to detect a small to medium-sized effect (f2 = .10) in a linear multiple regression with a power of 1-*β* = .80.In total, *N* = 136 participants completed the study, however *n* = 8 needed to be excluded due to technical failure additional to *n* = 5 due to lack of game structure comprehension, resulting in a final sample of *N* = 123 (65 female).[[5]](#footnote-5) Participants had a mean age of 22.7 years (*SD* = 5.0; *Min* = 17; *Max* = 60).[[6]](#footnote-6)

**Results and Discussion**

On a descriptive level, we observed a moderate level of punishment, about a third of participants (30.08%) punished in the first round, noticing a very solid level of cooperation (*M* = 258.98, *SD* = 152.55).[[7]](#footnote-7) Punishment investments refer to the invested resources (in Cent) and were set zero if no punishment was conducted, resulting in an average investment level of *M* = 16.71 Cents (*SD* = 35.36, *Min* = 0, *Max* = 200) and a point-biserial correlation with the binary punishment decision of *r* = .71, *p* < .001 (for an overview of descriptive statistics and bivariate correlations of the main variables see Table 1).

In line with our hypothesis (H1), we find a main effect of spontaneous punishment, significant for the continuous punishment investments (OLS regression: *β* = -.23, *p* = .01; see Figure 2) and a tendency in the binary decision (logistic regression: odds ratio = .09, *p* = .07). Unexpectedly, there was no correlation between punishment and SVO (*r* = -.01, *p* = .91), though a moderate effect for people who are inequality averse to punish more frequently than joint gain maximizer (JGM) (*r* = .25, *p* = .075). On theoretical grounds, this is meaningful as for joint gain maximizers it is not justifiable to punish as this minimizes the total welfare for the group. Inequality averse persons, however, accept a lower joint gain for the sake of an equal distribution. Contrary to our expectation (H2), spontaneous punishment is unlike spontaneous cooperation not driven by SVO (neither the binary punishment: odds ratio = .99, *p* = .90, nor the analysis of continuous punishment investments: *β* = -.09, *p* = .36). However, situational above-average cooperation has a strong influence on punishment behavior: above-average contributors are more likely to punish than participants who cooperated below-average (*r* = .35, *p* < .001, see Table 1). As for the interaction with decision time, there is a trend for the continuous punishment investments (*β* = -.18, *p* = .07, one-sided) though not reflected in the binary punishment (odds ratio = .18, *p* = .43). As on a conventional significant level reflected in the even higher powered overall analysis (see results section of the overall-analysis after Study 2) above-average contributors start to punish severely, but decrease in their punishment behavior over the course of decision time. Below-average contributors invest constantly few resources for punishment (see Figure 3).

Testing the influence of negative affect (subscale upset) on punishment, there are distinct correlations indicating increasing punishment with negative affect (in the binary as well as continuous punishment decisions: *r* = .41 and *r* = .38, respectively, both *p* < .001). Concerning the moderating influence of negative affect on spontaneous punishment (H3), we again ran an OLS regression for the continuous punishment investments and find no conditional effect of negative affect on spontaneous punishment (punishment investments: *β* = .02, *p* = .86). However, there is a trend in the unexpected direction for the binary punishment decision (odds ratio = 2.3, *p* = .087). The question remains whether our affect measurement after the second round might have inflated the results for the first round. As we measured the PANAS after the second PGG round, we might have gained an affect that refers to the more recent interaction, i.e., the second round or at least consist of an affect aggregation of both rounds, implicating also a potential affect regulation of the first round. To meet this criticism and to disentangle the motivation to punish of future-oriented deterrence and the retributive character, we run a one-shot PGG in Study 2.

**Study 2**

As indicated, the iterative PGG in Study 1 does not allow for a differentiation in the motivations of punishment as the results entail next to the retributive character also a rational component to deter from further freeriding in the second round. In order to disentangle the influence of that rational motive, the second study consisted of a one-shot PGG.

**Method**

*Materials & Procedure.* We took the materials of study 1, only removed the second round in the PGG and added one item for further clarification on the underlying motivations at the very end of the study (see appendix).

*Participants.* Anapriori power analysis using G\*Power (Faul et al., 2009) revealed a required sample size of *N* = 143 to detect a small to medium-sized effect (f2 = .10) in a linear multiple regression with an increased number of predictors in order to test with sufficient power (1-*β* = .80) for the interactions.We collected data of *N* = 168, but lost data of *n* = 14 due to lack of game comprehension or technical failure, resulting in a final sample of *N* = 154.[[8]](#footnote-8) Participants who were again invited via the ORSEE database (Greiner, 2004) of the Max Planck Institute for Research on Collective Goods in Bonn, Germany. Their mean age consist of 23.29 years (SD = 5.20) and 78 females[[9]](#footnote-9).

**Results and Discussion**

Again, we find in line with our hypothesis (H1) a main effect of spontaneous punishment, significant for the continuous punishment investments (OLS regression: *β* = -.19, *p* = .02; see Figure 2), though not for the binary decision (logistic regression: odds ratio = .76, *p* = .78). Similar to Study 1, spontaneous punishment is unlike spontaneous cooperation not driven by SVO (H2, neither for the binary punishment: odds ratio = .98, *p* = .80, nor for the analysis of continuous punishment investments: *β* = .01 *p* = .93). However, situational above-average contributions have a strong influence on punishment behavior, both as a main effect (*r* = .40, *p* < .001, see Table 1) and also in interaction with decision time (punishment investments: *β* = -.28, *p* = .03, binary punishment: odds ratio = .00, *p* = .003, see Figure 3): For above-average contributors, there are spontaneously conducted high punishment investments that decrease over time. In contrast, below-average contributors show a minimal punishment investment level that is stable over the course of decision time.

In line with H3, spontaneous punishment is strongly conditional on negative affect (punishment investments: *β* = -.43, *p* < .001, binary punishment: odds ratio = .28, *p* = .005) and in comparison to study 1 also in the expected direction in that there is a spontaneous tendency for highly upset people to punish that decreases over time. Less upset persons show as expected a generally limited willingness to punish their team members that is not changing over the course of decision time.

**Overall-analysis**

As or studies were very similar with the only distinction of an iterative vs. one-shot PGG, we conduct an overall analysis in order to verify our results with the highest power possible.[[10]](#footnote-10) Overall, there is no spontaneous punishment (H1) for the binary punishment decision (logistic regression: odds ratio = .34, *p* = .17), however, when differentiating the amount of punishment (no punishment = 0 Cent Investment), a continuous decrease of punishment investments turn visible (OLS regression: *β* = - .20, *p* = .001).

As the separate analyses of both studies already suggest, spontaneous punishment is – unlike spontaneous cooperation – not driven by prosocials (H2), neither for the binary punishment decision (odds ratio = .87, *p* = .97), nor for the continuous punishment investment (*β =* -.03*, p* = .63)*.* Interestingly, there is even no main effect of SVO on Punishment (*r* = -.07, *p* =.23) nor for Inequality Aversion - as a facet of prosociality - and punishment behavior (*r* = -.05, *p* = .625).[[11]](#footnote-11) In line with Fehr and Gächter (2002), however, who identified situational above-average cooperation as one source of punishment, spontaneous punishment decisions are driven by above average contributors (binary punishment: odds ratio = .02, *p* = .02; punishment investments: *β* = - .23, *p* = .01). However, in our studies contributions are only moderately related to dispositional SVO (*r* = .19, *p* = .02).

Overall, our data suggests that spontaneous punishment decisions are driven by negative affect, only marginal significant for binary punishment decisions (odds ratio = .60, *p* = .07), however highly significant for the continuous punishment investment (*β =* -.26, *p <* .001). When exploring the relation between negative affect, above-average contributions and decision time, we find a significant three-way interaction in that spontaneous punishment is only valid for above-average, highly upset contributors (binary punishment: odds ratio = .32, *p* = .09; see Table 3; continuous punishment: *β =* -.25, *p* = .01, see Table 4 & Figure 5). However, our correlative design does not allow for any causal conclusions whether the above-average experience the negative affect as a result of having been exploited although the order of measurement suggests it and the instructions of the PANAS explicitly requested to consider one’s feelings towards the group members’ contributions. As negative affect is overall moderately correlated with above-average contributions (*r* = .22, *p* < .001), we also test for a moderated mediation model and but do not find evidence in favor of a mediating role of negative affect for the interaction of above-average contributions and spontaneous punishment (average causal mediated effect = -3.94, 95% CI [-13.69, 3.72]).[[12]](#footnote-12)

**Replicational results of spontaneous cooperation for prosocials**

Our design allows a direct replication analysis of the findings of Mischkowski and Glöckner (2016), with the only design modification of an anticipated punishment opportunity. In both studies separately as well as in an overall analysis, we replicate the spontaneous cooperation effect for prosocials and overall also the main effect of spontaneous cooperation. The interaction of response time and SVO follows the presented pattern in that spontaneous cooperation behavior is only valid for prosocial, but not for proselfs (see Table 5). Independent of decision time, their cooperation behavior is comparably low. Using the extended SVO Slider Measure, we are able to further differentiate between JGM and IA persons and find in both studies a significant interaction of response time and IA on cooperation behavior (Tobit regression: *b* = -1464.09, *t*(109) = -2.45, *p* = .02). For JGM the cooperation behavior is consistently *high*, but IA persons reduce their cooperation over time. Although this analysis has been exploratory, it is theoretically meaningful as JGM have a rational interest in contributing as this maximizes the joint outcome. For IA persons, deliberation might put in their focus that cooperation exposes them to potential exploitation and therefore increasing risk of inequality. These results need to be further explored (especially without anticipated punishment), but may provide a starting point to further investigate differences among prosocials.

**General Discussion**

There are several plausible assumptions why punishment should be conducted spontaneously – as being a highly aroused, affect-driven behavior as well as it might be driven by fairness concerns that are expressed automatically. To test these different moderators - that are not mutually exclusive - we run two studies of one iterative and an one-shot public goods game, both including a second-party punishment option. We measured response time of cooperation and punishment decisions and measured affect after punishment decision using the PANAS scale. We find in both studies that upset as the subscale of negative affect moderates spontaneous punishment in that punishment decisions are quicker for persons upset about the contribution behavior of their group members. Overall, there is no main effect of spontaneous punishment in the binary punishment decision. No interaction is found with dispositional social preferences, i.e., SVO or inequality aversion, but rather with situational above-average contributions. Spontaneous punishment behavior therefore seems to be rather driven by its situational, retributive character – at least for a second-party setting.

*Limitations.* The question remains whether altruistic punishment (third-party punishment) offers a more suitable setting for social preferences to exert influence. First of all, this might be as one person in the role of the punisher might have a different perspective on the wealth distribution for the whole group, less than an involved individual, trying to end up best.[[13]](#footnote-13) Furthermore, moral values might be stronger activated in third-party punishment situations, which might lead to a more pronounced relation with inequality aversion. Still, also in third party punishment emotions and above-average contributions (Fehr & Fischbacher, 2004; Fehr et al., 2002; Fehr & Gächter, 2002) are important factors driving punishment. We replicate these results for second-party punishment decisions and – even more importantly – show their influence over the course of decision time.

Note that our design does not allow shedding light on the underlying processes, what drives the increased punishment investments, e.g. if one ‘makes use of the occasion’ and decides on short-notice to additionally punish another member than originally considered when making the binary punishment decision. We can also only speculate which feedback information (e.g., contribution level or total PGG earnings) were decisive for or against punishment. Further research could you processing technologies (i.e., eye- or mouse-tracking) in order to shed light on the underlying cognitive processes.

Recently, a debate has arisen whether response time allow for a reverse induction of cognitive processes (Krajbich, Bartling, Hare, & Fehr, 2015), but might better reflect the extent of decision conflict (Kieslich & Hilbig, 2014). We do not argue that shorter response times necessary reflect intuitive processing. Rather, in line with a default-interventionist model (Horstmann, Ahlgrimm, & Glöckner, 2009) they provide the individual default which can be intervened by longer deliberation. The conclusion therefore persists that second-party punishment is not a spontaneously conducted act by prosocials, but rather a (negative) affect-driven response to subjectively perceived injustice. The different factors underlying punishment point out the different motivations that might drive punishment decisions and also show that punishment is a much more complex phenomenon than cooperation behavior is. From a normative perspective, cooperation is clearly advantageous for the group. In contrast, punishment behavior might be negatively connoted as consists by all means of a destruction of resources and therefore implies a loss for the punisher as well as the punished one.

**Conclusion**

Similar to a spontaneous cooperation effect, we find that the invested resources to punish decrease with increasing decision time. There is no main effect of spontaneous punishment behavior for the binary punishment decision, though. Even though dispositional social preferences are not moderating spontaneous punishment, situational above average cooperation behavior does promote spontaneous punishment. Exploratory analyses of moderated mediation models lead to the tentative conclusion that negative affect can explain for the increased punishment of above-average contributors. Punishment as a second order public good therefore seems to follow a different motivation than cooperation behavior, even though the mere behavior follows a similar pattern

Our studies lead to the tentative conclusion that punishment as a second order PGG differ from first order cooperation behavior as its spontaneity is conditional on a high negative arousal. Interestingly, the perception of punishment differs between (non-)punishers being fair and just for both decisions. Further research is needed to extend the results to third party punishment (altruistic punishment) in order to test whether prosocial traits play a more important role when there is no diffusion of responsibility who is in the role to establish fairness and no personal involvement within the group setting.

**Tables**

Table 1. *Means, Standard Deviations (in Parentheses) and Inter-Correlations in Study 1*.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Scale | Mean (SD) | Correlations |
|  |  |  | 1.  | 2.  | 3. | 4.  | 5.  | 6.  | 7.  | 8.  |
| 1. Punishment | 0,1  | 30.08% | 1 |  |  |  |  |  |  |  |
| 2. Punishment Investments  | ≥ 0, in Cents | 16.71 (35.63) | .71\*\*\*(123) | 1 |  |  |  |  |  |  |
| 3. Contributions | 0-400 Cents | 258.98 (152.55) | .30\*\*\*(123) | .28\*\*(123) | 1 |  |  |  |  |  |
| 4. Above-average Contributions | Dummy coded (0,1) | 51.22% | .35\*\*\*(123) | .32\*\*\*(123) | .77\*\*\*(123) | 1 |  |  |  |  |
| 5. SVO angle | In degree (+/- 180)  | 22.97 (13.75) | -.01(117) | .03(117) | .23\*(117) | .06(117) | 1 |  |  |  |
| 6. Inequality Aversion | 0-1 (0 = IA, 1 = JGM) | .25 (.26) | -.24+(54) | -.16(54) | -.01(54) | -.21(54) | -.02(54) | 1 |  |  |
| 7. Negative Affect (Upset) | 0-12 | 3.61 (3.64) | .41\*\*\*(123) | .38\*\*\*(123) | .02 (123) | .12(123) | -.20\*(117) | -.00 (54) | .92(123) |  |
| 8. RT Pun. (RT in log10(sec)) | Log 10 RT | .86 (.16) | -.17+(123) | -.23\*(123) | -.09(123) | -.16+ (123) | -.02(117) | -.24\*\*\*(123) | .01 (54) | 1 |

*Note*. + *p* < .1, \* *p* < .05, \*\* *p* < .01, \*\*\* *p* < .001 (two-sided). SVO = Social Value Orientation. Note that results only refer to the first round of the PGG. For binary variables (above-average contributions and punishment) we report point-biserial correlations. Numbers of observations are given in parentheses. Cronbach’s alpha is provided in the diagonal for negative affect.

Table 2. *Means, Standard Deviations (in Parentheses) and Inter-Correlations in Study 2*.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Scale | Mean (SD) | Correlations |
|  |  |  | 1. | 2.  | 3. | 4. | 5. | 6.  | 7.  | 8. | 9.  |
| 1. Punishment | 0,1  | 59.09% | 1 |  |  |  |  |  |  |  |  |
| 2. Punishment Investments  | >= 0, in Cents | 25.35 (47.12) | .65\*\*\*(154)  | 1 |  |  |  |  |  |  |  |
| 3. Contributions | 0-400 Cents | 262.12 (156.14) | .28\*\*\*(154) | .26\*\* (154) | 1 |  |  |  |  |  |  |
| 4. Above-average Contributions  | Dummy coded (0,1) | 50.65% | .40\*\*\*(154) | .34\*\*\*(154) | .71\*\*\*(154) | 1 |  |  |  |  |  |
| 5. SVO angle | In degree (+/- 180)  | 23.81 (15.42) | -.12 (147) | -.03(147) | .15+(147) | .03 (147) | 1 |  |  |  |  |
| 6. Inequality Aversion | 0-1 | .30 (.32) | .08(58) | .02(60) | .01 (60) | .05(60) | .30\*(60) | 1 |  |  |  |
| 7. Negative Affect (Upset) | 0-12 | 2.72 (3.30) | .20\*(153) | .21\*(153) | .16+(153) | .30\*\*\* (153) | -.14+  (146) | .07 (60) | .92(153) |  |  |
| 8. RT Punishment (RT in log10(sec)) | Log 10 RT | .90 (.17) | -.02(154) | -.19\*(154) | .05(154) | .02(154) | .04(147) | .23+ (60) | -.13(153) | 1 |  |
| 9. RT Pun. Investments (in log10(sec)) | Log 10 RT | 1.62 (.23) |  | .18 (63) | .00 (63) | -.07(63) | -.08(60) | -.27(20) | -.08(63) | .09(63) | 1(63) |

*Note*. + *p* < .1, \* *p* < .05, \*\* *p* < .01, \*\*\* *p* < .001 (two-sided). SVO = Social Value Orientation. For binary variables (above-average contributions and punishment) we report point-biserial correlations. Numbers of observations are given in parentheses. Cronbach’s alpha is provided in the diagonal for negative affect.

Table 3. Logistic regression (reported in odds ratios) of decision times, negative affect (upset), above-average contributions on punishment decisions

|  |  |  |  |
| --- | --- | --- | --- |
| Punishment (binary choice) | Overall Analysis | Study 1 | Study 2 |
| Decision Time (DT in log10(sec)) | 11.85(1.52) | .10(-.80) | 682.52\*(2.03) |
| Above-average contributions | 4.62\*\*\*(5.02) | 4.34\*\*(2.74) | 6.48\*\*\*(4.15) |
| Negative Affect (Upset)  | 1.13(1.64) | 1.23(1.59) | 1.06(.51) |
| Interaction of DT \* above-average | .01\*(-2.38) | 5.15(.47) | .00\*\*(-2.98) |
| Interaction of DT \* upset | 1.23(.38) | 3.56(1.42) | 1.81(.48) |
| Interaction of above-average\* upset | 1.02(.22) | 1.22(1.11) | .95(-.35) |
| Interaction of DT \* above-average \* upset | .32+(-1.69) | .39(-.78) | .13(-1.49) |
| Constant | .14\*\*\*(-6.37) | .14\*\*\*(-4.49) | .23\*\*\*(-3.99) |
| Observations | 276 | 123 | 153 |
| Pseudo R2 | .20 | .28 | .21 |

+ p < .1, \* p < .05, \*\* p < .01, \*\*\* p < .001 (two-sided)

Note. z-values are presented in parentheses; continuous predictors are centered. Overall analysis also includes a study dummy which is not reported.

Table 4. OLS regressions of decision times, negative affect (upset), above-average contributions on punishment investments

|  |  |  |  |
| --- | --- | --- | --- |
| Invested Punishment Resources (in Cent) | Overall Analysis | Study 1 | Study 2 |
| Decision Time (DT in log10(sec)) | -3.79 (-.17) | -7.04 (-.27) | -9.81(-.21) |
| Above-average contributions | 21.12\*\*\* (4.55) | 14.89\*(2.54) | 27.48\*\*\*(3.82) |
| Negative Affect (Upset)  | .34 (.31) | 1.04(.73) | -.29(-.16) |
| Interaction of DT \* above-average | -67.92\* (-2.37) | -36.13 (-.99) | -97.37+(-1.84) |
| Interaction of DT \* upset | -2.83 (-.37) | 4.59 (.55) | -10.90(-.59) |
| Interaction of upset \* above-average | 1.37 (.94) | 3.73+(1.93) | .68(.30) |
| Interaction of DT \* above-average \* upset | -24.19\* (-2.59) | -8.14(.66) | -26.35(-1.33) |
| Constant | 1.038 (.25) | 5.82(1.41) | 9.24+(1.72) |
| Observations | 276 | 123 | 153 |
| Adjusted R2 | .25 | .25 | .28 |

+ p < .1, \* p < .05, \*\* p < .01, \*\*\* p < .001 (two-sided)

Note. t-values are presented in parentheses; continuous predictors are centered. Overall analysis also includes a study dummy which is not reported.

Table 5. Tobit regressions of decision times and Social Value Orientation on contributions

|  |  |  |  |
| --- | --- | --- | --- |
| Contribution (in %) | Overall Analysis | Study 1 | Study 2 |
| Decision Time (DT in log10(sec)) | -272.94\*\*  (-3.15) | -204.73+ (-1.69) | -379.58\*\*  (-3.00) |
| Social Value Orientation (SVO angle in degree) | 4.51\*\*(3.19) | 6.22\*(2.57) | 3.30+ (1.93) |
| Interaction of DT \* SVO | -18.70\*\* (-2.99) | -21.07\* (-2.12) | -17.35\*  (-2.19) |
| Constant | 370.53\*\*\* (11.45) | 274.07\*\*\* (10.34) | 359.94\*\*\* (12.60) |
| Observations | 264 | 117 | 147 |
| Pseudo R2 | .01 | .02 | .02 |

+ p < .1, \* p < .05, \*\* p < .01, \*\*\* p < .001 (two-sided)

Note. t-values are presented in parentheses; predictors are centered. Overall analysis also includes a study dummy which is not reported.

**Figures**

****

*Figure 1*. Procedure of both studies. Note that study 1 consisted of two rounds of the PGG.



*Figure 2*. Spontaneous Punishment (overall analysis).



*Figure 3*. Spontaneous Punishment for above-average contributors (overall analysis).



*Figure 4*. Spontaneous Punishment for upset persons (overall analysis).



*Figure 5.* Three-way interaction of spontaneous punishment for above-average contributors that are highly upset (overall analysis).

**References**

Ackermann, K. A., & Murphy, R. O. (2012). Tutorial on how to evaluate the SVO Slider Measure’s secondary items. from ttp://vlab.ethz.ch/svo/SVO\_Slider/SVO\_Slider\_scoring.html

Balliet, D., Parks, C., & Joireman, J. (2009). Social value orientation and cooperation in social dilemmas: A meta-analysis. *Group. Process. Intergroup. Relat., 12*(4), 533-547. doi: 10.1177/1368430209105040

Bogaert, S., Boone, C., & Declerck, C. (2008). Social value orientation and cooperation in social dilemmas: A review and conceptual model. *British Journal of Social Psychology, 47*, 453–480. doi: 10.1348/014466607X244970

De Dreu, C. K., Greer, L. L., Handgraaf, M. J., Shalvi, S., Van Kleef, G. A., Baas, M., . . . Feith, S. W. (2010). The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science, 328*(5984), 1408-1411.

De Quervain, D. J., Fischbacher, U., Treyer, V., & Schellhammer, M. (2004). The neural basis of altruistic punishment. *Science, 305*(5688), 1254.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G\* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149-1160.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and human behavior, 25*(2), 63-87.

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong Reciprocity, Human Cooperation and the Enforcement of Social Norms. *Human Nature, 13*, 1-25. doi: 10.1007/s12110-002-1012-7

Fehr, E., & Gächter, S. (2002). Altruistic Punishment in Humans. *Nature, 415*, 137-140. doi: 10.1038/415137a

Fiedler, S., Glöckner, A., Nicklisch, A., & Dickert, S. (2013). Social Value Orientation and information search in social dilemmas: An eye-tracking analysis. *Organizational Behavior and Human Decision Processes, 120*(2), 272-284. doi: 10.1016/j.obhdp.2012.07.002

Greiner, B. (2004). An Online Recruitment System for Economic Experiments. In K. Kremer & V. Macho (Eds.), *Forschung und wissenschaftliches Rechnen 2003* (pp. 79-93): Göttingen: Ges. für Wiss. Datenverarbeitung.

Gummerum, M., Van Dillen, L. F., Van Dijk, E., & López-Pérez, B. (forthcoming). Costly third-party interventions: The role of incidental anger and attention focus in punishment of the perpetrator and compensation of the victim. *Journal of Experimental Social Psychology*.

Hardin, G. (1968). The tragedy of the commons. *Science, 162*(3859), 1243-1248. doi: 10.1126/science.162.3859.1243

Horstmann, N., Ahlgrimm, A., & Glöckner, A. (2009). How distinct are intuition and deliberation? An eye-tracking analysis of instruction-induced decision modes. *Judgment and decision making, 4*(5), 335–354.

Kieslich, P. J., & Hilbig, B. E. (2014). Cognitive conflict in social dilemmas: An analysis of response dynamics. *Judgment and decision making, 9*(6), 510-522.

Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature communications, 6*.

Krohne, H. W., Egloff, B., Kohlmann, C.-W., & Tausch, A. (1996). Untersuchungen mit einer deutschen Version der "Positive and Negative Affect Schedule" (PANAS). *Diagnostica, 42*(2), 139-156.

Lieberman, M. D., Eisenberger, N. I., Crockett, M. J., Tom, S. M., Pfeifer, J. H., & Way, B. M. (2007). Putting feelings into words affect labeling disrupts amygdala activity in response to affective stimuli. *Psychological Science, 18*(5), 421-428.

Mischkowski, D., & Glöckner, A. (2016). Spontaneous cooperation for prosocials, but not for proselfs: Social value orientation moderates spontaneous cooperation behavior. *Scientific Reports, 6*.

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. J. (2011). Measuring Social Value Orientation. *Judgm. Decis. Mak., 6*(8), 771–781. doi: 10.2139/ssrn.1804189

Nielsen, U. H., Tyran, J.-R., & Wengström, E. (2014). Second thoughts on free riding. *Economics Letters, 122*(2), 136-139.

Peters, E., & Slovic, P. (2007). Affective asynchrony and the measurement of the affective attitude component. *Cognition and Emotion, 21*(2), 300-329.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature, 489*, 427-430. doi: 10.1038/nature11467

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nat. Commun., 5*, 1-12. doi: 101038/ncomms4677

Renkewitz, F., Fuchs, H. M., & Fiedler, S. (2011). Is there evidence of publication biases in JDM research? *Judgment and decision making, 6*(8), 870-888.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science, 300*(5626), 1755-1758.

Seithe, M., Morina, J., & Glöckner, A. (2015). Bonn eXperimental System (BoXS): An open-source platform for interactive experiments in psychology and economics. *Behavior Research Methods*, 1-22.

Tinghög, G., Andersson, D., Bonn, C., Böttiger, H., Josephson, C., Lundgren, G., . . . Johannesson, M. (2013). Intuition and cooperation reconsidered. *Nature, 498*, 427–430 doi: 10.1038/nature12194

Van Lange, P. A. (2000). Beyond self-interest: A set of propositions relevant to interpersonal orientations. *European review of social psychology, 11*(1), 297-331.

Verkoeijen, P., & Bouwmeester, S. (2014). Does intuition cause cooperation. *PloS One, 9*(5), e96654. doi: doi:10.1371/journal.pone.0096654

1. Related emotions refer to the antisocial component of punishment, i.e., the pleasure to punish those who contributed above-average. This motivation is a special case, with large theoretical and empirical grounds, but usually not the dominant strategy. Therefore it is not the main focus of this paper. [↑](#footnote-ref-1)
2. We do again not refer to antisocial punishment here which obviously enhances inequality, but is at the same time unlikely to be conducted by prosocials. [↑](#footnote-ref-2)
3. As we imply punishment having a high arousal, we refer to the subscale “upset” of the negative affect scale (see Circumplex Model of Affect, Russel & Feldman Barrett, 1999). [↑](#footnote-ref-3)
4. Control questions asked for the amount of contributions that maximize the personal vs. the group’s outcome as well as the necessary punishment investment in order to withdraw 1€ (approximately $1.10) from another group member. [↑](#footnote-ref-4)
5. Note that results do not change when including these participants. [↑](#footnote-ref-5)
6. Note that due to the collection of demographics in the online part, we have missing values for those participants who could not be matched, resulting in subsample for demographics of *N* = 117. [↑](#footnote-ref-6)
7. Note that we only refer to the first round of PGG without any learning or reputation effects. [↑](#footnote-ref-7)
8. Note that results don’t change when including the persons with lack of game comprehension. [↑](#footnote-ref-8)
9. Note that demographics were again collected online before the lab session. *n* = 7 participants of the lab sessions could not be matched, which is why there are missing values for their demographics and SVO. [↑](#footnote-ref-9)
10. Note we included a study dummy in all overall analyses. [↑](#footnote-ref-10)
11. Neither does Inequality Aversion - in contrast to JGM as subtypes of prosociality - moderate spontaneous punishment (*β* = - .03, *p* = .76; odds ratio = 1.54, *p* = .30). Note, however, that the power of this analysis is reduced as IA can only be meaningful interpreted for prosocials (Ackermann & Murphy, 2012). In our first study we had a sample of *n* = 54 prosocials for whom we could calculate an IA index between 0 and 1; zero indicating a perfect preference for IA. For our second study, there were *n* = 60 prosocials, resulting in an *N* = 114 for testing the moderating influence of IA on spontaneous punishment. [↑](#footnote-ref-11)
12. Recently, some interesting causal effects have been shown for incidental emotions in that they require attentional resources that lead to increased third-party punishment (Gummerum, Van Dillen, Van Dijk, & López-Pérez, forthcoming). We, however, focus at the heart of the process and hence measure integral emotions. [↑](#footnote-ref-12)
13. Several participants indicated in the free message form why they chose (not) to punish that they wanted to systematically weaken their team members in order to end up with the highest outcome of the group („Winning“). [↑](#footnote-ref-13)