Preprints of the
Max Planck Institute for
Research on Collective Goods
Bonn 2017/15

Abuse of Power
An experimental investigation of the effects of power and transparency on centralized punishment

Leonard Hoeft
Wladislaw Mill

MAX PLANCK SOCIETY

Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2017/15

Abuse of Power

An experimental investigation of the effects of power and transparency on centralized punishment

Leonard Hoeft / Wladislaw Mill

August 2017

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

Abuse of Power
An experimental investigation of the effects of power and transparency on centralized punishment

Leonard Hoeft[a,b], Wladislaw Mill[a,c,*]

[a] *International Max Planck Research School on Adapting Behavior in a Fundamentally Uncertain World.*
[b] *Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Straße 10, 53113 Bonn Germany.*
[c] *School of Economics and Business Administration, University of Jena, Bachstraße 18k, 07743 Jena, Germany.*

Abstract

We investigate power abuse of a single punisher in a public-goods-game subject to variations in punishment power and contribution transparency. We find a high amount of abuse across all conditions. More power led to more abuse over time, while transparency could only curb abuse in the high power conditions. These findings highlight the dangers of power centralization, but suggest a more complex relation of power and transparency

Keywords: Punishment; Public-Goods-Game; Designated Punishment; Abuse; Transparency; Power
JEL: H41, C92, K42

Evil is nourished and grows by concealment.

Virgil

1. Introduction

In April 2016, the Panama Papers revealed how wealthy individuals and public officials were able to employ (mostly legal) strategies of tax evasion, systematically avoiding their share of contribution to the public-goods. This was largely interpreted as abuse of power, defined as the improper use of (political) power for illegitimate private gain. This restarted an old debate on the relationship between power and transparency: Does transparency prevent corruption and abuse of power? Laboratory experiments have a long history investigating the relationship between contributions to the public-goods and punishment power, which mostly painted an optimistic story: Participants selflessly use (peer) punishment to solve social dilemmas (Fehr and Gächter, 2002), even if they are a centralized second- or third-party punishers (O'Gorman et al., 2009, Fehr and Fischbacher, 2004, Baldassarri and Grossman, 2011a). So far, the second-party punisher was randomly determined each round, effectively prohibiting abuse as no participant could systematically use his position of power to exempt himself from the enforced contributions. This leaves open one important aspect of power centralization in a community: Is there a corrupting effect in absence of checks and balances? How do the privileged few use their power? These power discrepancies are ubiquitous: From managers to legal officials, police or politicians, people are in positions to enforce beneficial rules while not adhering themselves. Consequently, fear of self-serving use of power by the elites and authorities has been a common theme of diverse societies and organizations.

*Corresponding author; Tel. +493641 930410; Bachstraße 18k, 07743, Jena Germany.
Email addresses: hoeft@coll.mpg.de (Leonard Hoeft), wladislaw.mill@uni-jena.de (Wladislaw Mill)

We close this gap in the literature on centralized second-order punishment and investigate institutional abuse under varying amounts of power (measured in available punishment points) and transparency (understood as contribution transparency). We find a large amount of power abuse robust to different definitions of abusive behavior. Punishers enforce contribution norms they significantly undercut themselves. Without transparency, higher power leads to higher imposed norms and more abuse over time. Surprisingly, however, transparency only curbs the abuse in the high-power treatment.

2. Literature

Punishment has been one of the popular remedies for social dilemmas since the groundbreaking work showing that participants are willing to provide a second-order public-good of enforcement (Fehr and Gächter, 2002). In the laboratory, peer punishment can effectively prevent the eventual decline of the first-order public-good.

Peer punishment may, however, fall short of the desired goal. It can be highly inefficient when punishers fail to coordinate and therefore punish too little or too much. This is especially pertinent as each would prefer the other to bear the costs of punishing, which could lead to a failure to coordinate so that no punishment is meted out at all. Furthermore, the external validity of peer punishment remains uncertain (Guala, 2012). All but small scale, close-knit societies rely on some form of institutional punishment. Empirical studies could show that participants exhibit a preference for such institutions (Zhang et al., 2014, Traulsen et al., 2012, Hilbe et al., 2013), even being willing to pay for them (Andreoni and Gee, 2012). If institutions are available, participants can organize to implement them efficiently (Putterman et al., 2011) and their effectiveness depends on whether they are endogenously chosen (Markussen et al., 2014). This strain of literature models institutions as a centralized and automatic punishment mechanism: It is exogenously determined and fixed. In real societies, the institutions depend on members of society themselves and are therefore not only endogenously implemented, but depend on the strategic choices of their members.

To see how participants use institutional power, experiments investigated how subjects would fill roles of designated punishment. If only one person is endowed with punishment power in a public-goods-game as a second-party (O'Gorman et al., 2009) or third-party punisher(Baldassarri and Grossman, 2011b), the social dilemma is mitigated, even at a personal cost. This held even when strategic punishment was excluded, as the second-party punisher was randomly rematched every round and the third party punisher did not benefit from group contributions. Third-party punishers actively promote their own social preferences (Engel and Zhurakhovska, 2017). While no design targets the goals of single second-party punishers, some studies shed light on those of second- and third-party peer punishers with conflicting results: Carpenter and Matthews (2009) find that second-party punishers enforce absolute and fixed norms, while third-party punishers enforce norms relative to group contributions, yet in Carpenter and Matthews (2012) second-party punishers enforce conformism and third parties fixed minimum contribution norms. According to Zhou et al. (2017), third parties punish more frequently, severely and less antisocially. Motivations considered for deciding to punish reach from fairness norms (Falk et al., 2005), egalitarian motives (Johnson et al., 2009, Leibbrandt and López-Pérez, 2012), reputation gain and leadership (O'Gorman et al., 2009) to destructive impulses such as spite and retaliation (Herrmann et al., 2008, Houser and Xiao, 2010, Falk et al., 2005). Participants seem to understand that there are those who punish prosocially: Fehr and Williams (2013) show that subjects manage to elect cooperative leaders into positions of power, and cooperative non-punishers are willing to empower those who punish in the interest of the group (Gross et al., 2016). Although there are circumstances where decentralized peer punishment is preferred, under imperfect information third-party punishers (who, however, benefit from contributions) are favored (Nicklisch et al., 2016).

While the literature focuses on the positive externalities of punishment and the willingness to altruistically provide it, the dark sides of punishment have only recently attracted attention: Allowing for counter-punishment leads to unproductive revenge cycles (Nikiforakis and Normann, 2008), and sometimes punishment is directed at cooperators (Herrmann et al., 2008, Cinyabuguma et al., 2006). A more prevalent fear in modern societies, however, is not the vengeful use of power by citizens, but the abuse of power by those in sole positions of power, mainly the authorities. In a sender-receiver game, third-party punishers will punish senders even when they are honest if they can profit (Xiao, 2013), although this is reduced if they have to

provide justification (Xiao and Tan, 2014). In these cases, the norm-communicating function of punishment is impeded. Punishment could, however, be corrupted in more subtle and common ways: People in positions of power often used it to further their own agenda or to exempt themselves from duties they enforce onto others. Here, an important difference between second- and third-party punishment exists: While the third-party punisher can not profit from his own punishment, the second-party punisher can. Surprisingly, to the best of our knowledge no study has explored the behavior of a designated second-party punisher that is fixed over multiple rounds. This study contributes to the existing literature by exploring their punishment behavior, specifically focusing on power abuse. To exclude inequality, reputation, and leadership concerns, we use a setting in which only the punisher always gets feedback on individual contributions and can punish costlessly. To introduce different levels of power, we vary the amount of punishment points available.

In addition to the literature on punishment and its motivations, our study also contributes to informational feedback in relation to punishment. In our design, only the information of the non-punishers is varied: In the high information treatment, the non-punishers gain additional feedback on individual contributions, in the low information treatment only on aggregates. Absent of punishment, this leads to no contribution differences (Croson, 2001). Although studies suggest that if punishers themselves are subject to noise in feedback of others behavior the cooperation is more difficult to sustain (Patel et al., 2010, Ambrus and Greiner, 2012, Fischer et al., 2016) this does not necessarily shed light on how non-punishers react to informational differences. As transparency in public-goods with heterogeneous action spaces improves contributions only when punishment exists (Khadjavi et al., 2014), we might at least suspect that more information does not always benefit contributions when some inequality exists in society.

In our design, rational choice theory, assuming selfish and money-maximizing participants, would predict that every participant unequivocally defects and punishers do no use their punishment power. There is no immediate way to form predictions based on the previous experimental evidence, as the possibility of abuse we are interested in is only generated by our new design, in which a singular second-party punisher holds all punishment power. Therefore, we test straightforward and publicly held intuitions about the relationship between power and transparency.

We hypothesize that:

H1 *Higher power leads to more abuse.*
H2 *Transparency leads to less abuse.*
H3 *Higher power impedes the effect of transparency.*

The remainder of the paper is structured as follows: Section 3 will explain the design and the measurements of the experiment. In Section 4, we present the results. Section 5 concludes.

3. Materials and Methods

The main task of our experiment consisted of a public-goods-game task. We ran four treatments between subjects in a two-by-two design, varying high and low punishment power as well as high and low contribution transparency. Additionally, we elicited personality measurements such as SVO, spite, etc. We will first elaborate on the public-goods-game as the core of our experiment, then clarify the additional measurements and finally describe the data collection process.

3.1. Measurements

3.1.1. Public-goods-game task

In the public-goods task all participants were randomly assigned a role (punisher, non-punisher). They were also appointed to a group of four in which they remained for the duration of the public-goods-game (partner-matching)[1]. The public-goods-game was repeated for thirty rounds. Participants were instructed that each round would consist of three stages.

[1]For arguments for and against partner matching, see Andreoni and Croson (2008).

The first stage resembled a standard public-goods-game. Participants were asked to allocate 20 tokens to a private and public account (1 token = 25 euro cents). Tokens allocated to the private account were theirs to keep. Tokens allocated to the public account (c_i) had a marginal per-capita return (MPCR) of 0.5, so that each group member would receive 0.5 times the total contribution to the public-goods-game. The payoff π_i of the participant i can therefore be formalized in the following way:

$$\pi_i = 20 - c_i + 0.5 \cdot \sum_{j \in \{1,n\}} c_j \tag{1}$$

In the second stage, only the punisher (who was referred to as "D") would be informed about the first-stage contributions of all group members. The participants were shown in random order each round anew to rule out reputation effects from previous rounds. D (the punisher) was now asked to indicate how much he would punish subject i (ς_i, $i \neq D$)[2]. For this purpose he was equipped with 30 tokens in the "low" power treatment. In the high-power treatment, the punisher was equipped with 120 tokens. We set the low-power treatment to have enough punishment points to deter every participant from free-riding[3] to eliminate unobservable strategic considerations and focus on a purely behavioral effect. Each token could be used by the punisher to deduct a targeted subject's payoff by one token. Unused tokens were not added to the payoff of D to rule out equality concerns,[4] so the contributions of the punisher could be compared to the contributions of others directly. The other three group members were just shown a blank screen asking them to wait for the decision of the punisher. The payoff π_i of the participant $i \neq D$ can therefore be formalized in the following way (the payoff of the punisher is described by equation (1)):

$$\pi_i = 20 - c_i + 0.5 \cdot \sum_{j \in \{1,n\}} c_j - \varsigma_i \tag{2}$$

In the third stage (feedback stage), participants were informed about their own contribution to the private and group account, the overall group contribution, their own punishment (reduction), and their payoff. Non-punishers were informed of the contributions of other group members only in the high-transparency treatment. Non-punishers were never informed of punishment meted out to others - this was made public in the instructions to avoid leadership and reputational concerns.

Hence, we used a 2 (low power vs. high power) × 2 (low transparency vs. high transparency) between-subjects design. We will denote the high-power treatments (where the punisher was equipped with 120 punishment tokens) by HighPwr, the low-power treatments (where the punisher was equipped with 30 punishment tokens) by LowPwr. The high-transparency treatments (where all the contributions where public knowledge) will be denoted by HighTrans and the low-transparency treatments (where only the punishers knew individual contributions) by LowTrans.

3.1.2. Additional measurements

We also collected data on spite (Marcus et al., 2014), rivalry & narcissism (Back et al., 2013), and social value orientation (SVO) to increase the robustness of our results.

To measure SVO, we used the 6-items primary ring matching version of the Slider Measure (see Murphy et al., 2011, Murphy and Ackerman, 2014, for detailed implementation). At the end of the experiment, only one of the 6 items was randomly chosen to become payoff-relevant in case this task was paid. Either

[2]To avoid framing and demand effects, we referred to the act as "reducing the payoff".

[3]Note that the benefit of free-riding, compared to full contribution, is 10 tokens. If the punisher were confronted with three free-riders and utilized all 30 punishment tokens, he could make every free-rider indifferent between free-riding and fully contributing, by punishing each with 10 tokens. As soon as one subject contributes more than zero, the punisher can already make contributing a preferential option. Hence, 30 tokens are sufficient to ensure punishment to be deterrent.

[4]In case of payoff-relevant equipment, the punisher could contribute more in stage one, anticipating extra gains in the second stage. If there was no extra equipment, the punisher could contribute less in stage one, compensating his extra expenditure in stage two.

the slider-measure or the public-goods-game task was chosen with equal probability to be payoff-relevant,[5] while the three questionnaires (spite, rivalry, & narcissism) were not incentivized.

Only one of the thirty rounds was payoff-relevant in case the public-goods-game was drawn to be payoff-relevant for the respective subject.

3.2. Participants and Procedure

384 participants (53 % female) were recruited with the online registration software Hroot (Bock et al., 2014). The experiment was conducted at the BonnEconLab and consisted of 16 sessions each with 24 participants. The participants' age ranged from 16 to 57 years (Mdn = 23). Most students were bachelor students (Semester Mdn=5). The average earning was 14.54 € (including a 4 € show-up fee) and the experiment lasted 1.5 hours (including setting, video instructions, payoff etc.). All measurements were computerized with the experimental software z-Tree (Fischbacher, 2007).

Participants were randomly assigned to computer cubicles. They received video instructions separately and the opportunity to ask questions for each task in the experiment.[6] First, they were asked to complete SVO measurements. Then, they participated in a public-goods-game for 30 rounds. After that, they completed questionnaires (spite, rivalry, & narcissism) and filled in socio-demographics. At last, they were presented with their payoff information and received their payoff privately.

4. Data Analysis

The subsequent data analysis will be structured as follows: We will start by defining abuse as the core concept of our study. Our data will demonstrate high levels of abuse overall, with power having a corrupting effect under low transparency. Combining power with high transparency leads to significantly less abuse even compared with low-power treatments. We will show that abusive behavior is driven by increasing contribution norms that punishers enforce, while themselves undercutting their norm and the contributions of non-punishers.

4.1. Abuse of Punishment Power

Our main goal in this paper is to investigate whether punishers abuse their position of power. We define abuse as the deviation of the punisher's contribution from his own imposed contribution norm.[7] Hence, a punisher who imposes a norm of 18 in the current round, but contributes only 5, is behaving abusively. How abusively the punishers behave is described by the difference between his imposed norm in the current round and his contribution in the respective round. In this example the amount of abusive behavior would be 13.

Our definition of abusive behavior builds on two concepts: the contribution of the punisher and the imposed norm. The punisher's contribution is simply the amount the punisher contributed to the public-good. The imposed norm, on the other hand, is how much the punisher expects others to contribute. We assume punishment to be an expression of the punisher's violated expectations, which in turn are based on a contribution norm: We see in the sanction of contributions an indicator that the implicit norm was violated, and if a contribution is not sanctioned, we consider the contribution to fulfill the norm.[8] For example, if a contribution of 18 is punished and a contribution of 19 is not punished anymore, the imposed norm is at least 18. Hence, we define the highest contribution still punished as the lower bound of the contribution

[5] Hence, only one random problem was selected to become payoff-relevant which is the only incentive-compatible mechanism (see Azrieli et al., 2015, for a detailed argument).

[6] The video instruction with English subtitles can be found in the supplementary materials. An English version of the handout as well as screenshots of the experiment can also be found in the supplementary materials.

[7] In Appendix A.1, we also consider alternative definitions of abuse. We consider a simplistic approach, merely comparing the punishers' contribution with the average of the non-punishers. The derived results are virtually identical.

[8] Note that we consider already a small punishment as an indicator for norm violations, and hence, our definition of the imposed norm does not hinge on the punishment strength. For an analysis of punishment strength, we refer the reader to Appendix A.2.

norm.[9] Even though once-established norms are rarely abandoned, we consider all rounds where a punisher did not enforce an already established norm as not abusive to be conservative in our estimates.[10]

Over all treatments, 83.3 % of subjects abused their power at least once. Table 1 reports this percentage for each of the treatments. The table also reports the average percentage of rounds in which subjects behave abusively, given that they behave abusively at least once in the whole experiment. High transparency does curb abuse in the high-power treatment, $t(45.7)= 2.5$, $p= 0.02$; however, it has no significant effect on the average frequency of abusive behavior under low power $t(45.6)= 0.8$, $p\geq 0.05$. As can be seen in Figure 1, the discussed differences in abusive behavior are visible and are increasing over time.

Treatment	Abused at least once (in %)	How often abused, if abused (in %)	95% CI	Groups
LowPwr x LowTrans	88	61	[46,76]	24
LowPwr x HighTrans	83	53	[35,71]	24
HighPwr x LowTrans	83	77	[65,89]	24
HighPwr x HighTrans	79	45	[25,65]	24

Table 1: Descriptives of abuse

A linear mixed-effects model with fixed effects on treatments, time, the interaction of time and treatments, and a control for the last round, as well as subject-specific random effects is reported in Table 2. All results are robust to controls, including age, gender, SVO, spite, narcissism, rivalry, and the interaction of period and the mentioned measures.[11]

We can see that abusive behavior increases over time. Hence, subjects learn to abuse their power.

Under low transparency, power corrupts: Punishers in the high-power treatment abused their position more strongly over time. This effect supports our first hypothesis, namely that high power leads to more abuse.

Surprisingly, the effect of transparency was the opposite of our expectations (hypothesis two): under low power, it marginally increased abusive behavior. Hence, transparency was not only not helpful, it was actually harmful in the low-power treatment.

Concerning our third hypothesis, namely that high power impedes the effect of high transparency (in the sense that increased transparency will not have an effect on abusive behavior under high power), we find, remarkably the opposite effect. Transparency curbed abuse over time under high power.

Result 1 a 83.3 % of all punishers abused their power at least once.

Result 1 b Abuse increased over time.

Result 1 c Power corrupts under low transparency: The increase in abusive behavior over time was stronger under high power compared to low power.

Result 1 d Transparency has a marginally significant negative effect on abusive behavior in the low power setting.

Result 1 e Increased transparency reduces abusive behavior significantly in the high power setting.

[9] As in all treatments the punisher had enough punishment points to deter any contribution behavior, strategic or scarcity considerations can be excluded.

[10] Note that the results do not change by using a more lenient approach. The lenient approach could be easily defended as a once-established norm implies a thread-level and hence results in similar beliefs for non-punishers, even if the norm is not enforced on rare occasions.

[11] Note: We use a linear model as this seems sufficiently reasonable given the development over time. However, we also relax this assumption in Appendix B.3.1. For that purpose, we estimate a common loess abuse spline over rounds over all treatments. Using a Bayesian approach, we compare the curvity of the contribution over time of all the treatments. All results are fundamentally identical to the results reported in this section.

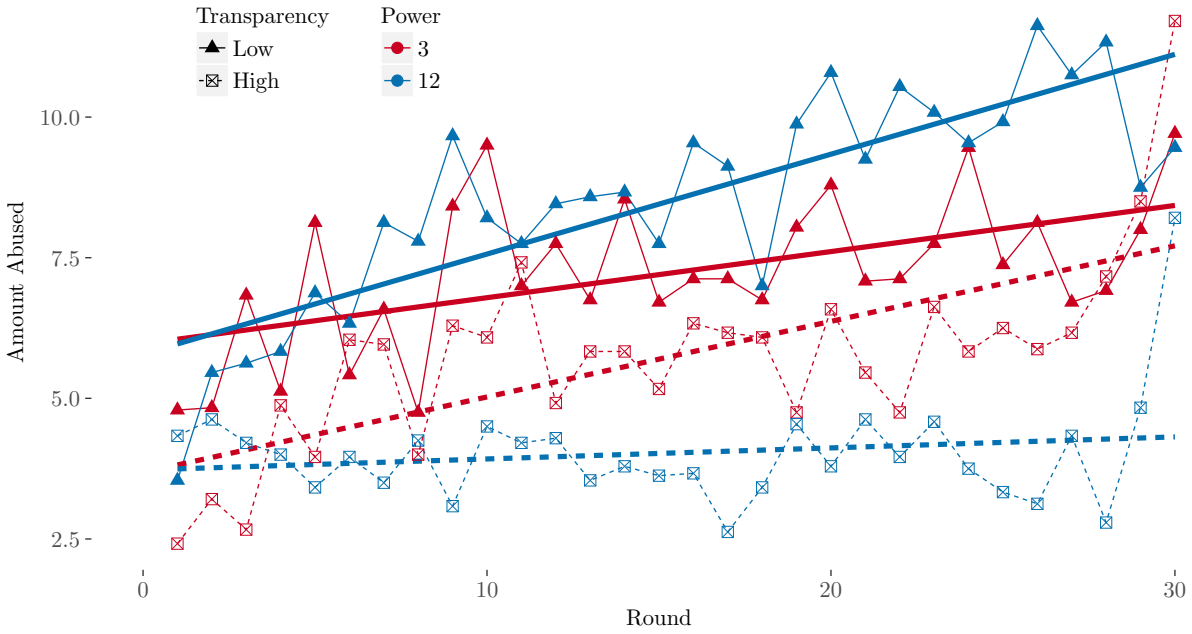

Figure 1: Amount abused in each of the four treatments over time with additional linear regression lines.
Blue lines represent the punisher's abusive behavior in the high-power treatments, while red lines represent the low-power treatments. The high-transparency treatments are represented by dashed lines with crossed cubes, while the low-transparency treatments are shown with solid lines and solid triangles. The thick lines denote the linear regression lines over time in each of the four treatments.

	Abuse			
	Abuse Amount		*High Power*	*Low Power*
Constant	6.11^{***} (1.24)	10.14^{**} (3.13)	5.88^{***} (1.23)	6.17^{***} (1.25)
HighTrans	-2.29 (1.75)	-2.64 (1.80)	-2.06 (1.74)	-2.29 (1.76)
HighPwr	-0.18 (1.75)	0.43 (1.78)		
HighTrans x HighPwr	0.22 (2.48)	-0.03 (2.51)		
t	0.07^{***} (0.02)	$0.08^{\cdot}$ (0.05)	0.17^{***} (0.02)	0.06^{**} (0.02)
t x HighTrans	$0.05^{\cdot}$ (0.03)	0.06^{*} (0.03)	-0.16^{***} (0.02)	$0.05^{\cdot}$ (0.03)
t x HighPwr	0.10^{***} (0.03)	0.09^{**} (0.03)		
t x HighTrans x HighPwr	-0.21^{***} (0.04)	-0.21^{***} (0.04)		
LastRound	2.15^{***} (0.48)	2.15^{***} (0.48)	1.28^{*} (0.61)	3.02^{***} (0.75)
Controls	×	✓	×	×
Observations	2,880	2,880	1,440	1,440
Log Likelihood	$-8{,}566.54$	$-8{,}589.98$	$-4{,}130.14$	$-4{,}406.59$
Akaike Inf. Crit.	17,155.09	17,225.95	8,274.28	8,827.17
Bayesian Inf. Crit.	17,220.71	17,363.16	8,311.19	8,864.08
Notes:				$^{\cdot}$p < 0.1;*p < 0.05;**p < 0.01;***p < 0.001;

Table 2: Mixed-effects model of the abusive behavior.
Controls include age, gender, spite, SVO, narcissism, rivalry, and the interaction of those variables with the period.

In the next two subsections, we will examine what was driving abusive behavior. For that purpose, we will first investigate how the imposed norm changes in the four treatments before describing the contribution behavior.

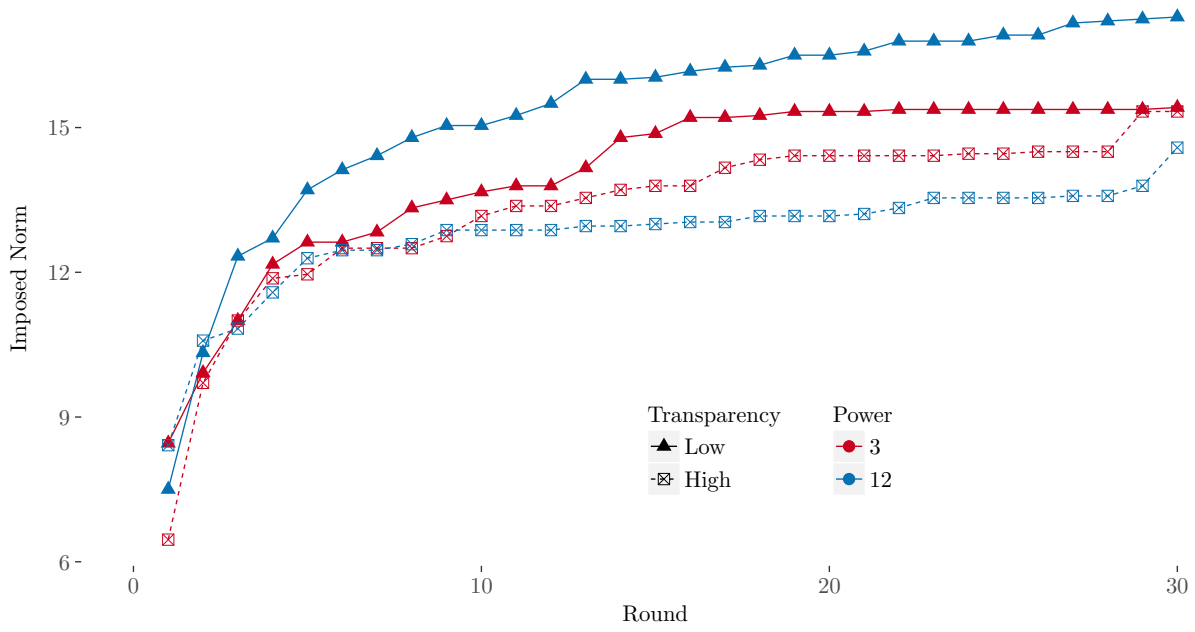

Figure 2: Imposed norms over time in each of the four treatments.
Blue lines represent the punisher's imposed norms in the high-power treatments, while red lines represent the low-power treatments. The high-transparency treatments are represented by dashed lines with crossed cubes, while the low-transparency treatments are shown with solid lines and solid triangles.

4.2. Imposed Norm

In this section, we investigate the norms punishers enforced.[12]

We measured for all punishers the norm they imposed. The average development over time of those norms by treatment are shown in Figure 2. By definition, the norm either stabilizes or increases. After roughly 10 rounds, the average imposed norm stabilizes and stays roughly constant for all treatments. Note that we found almost no instances of punishers ceasing punishment and giving up an already established norm.

Using a linear regression[13] for the imposed norm with fixed effects of treatments, time, the interaction of both and controlling for subject-specific random effects, we see in Table 3 that high power leads to a stronger increase in the imposed norm, although this effect is reversed if transparency is high.

All results are robust to the inclusion of controls.

Result 2 a High power leads to higher imposed norms over time. Transparency reverses this effect under high power.

Result 2 b Increased transparency does not affect the imposed norm under low power.

[12]Here, we look at the norms punishers imposed individually as we are interested in abusive behavior. For an analysis of punishment behavior on average see Appendix A.2

[13]Note: In Appendix B.2, we estimate a common loess spline over all treatments, and using a Bayesian approach, we compare the curvity of all the treatments. The implication of the results are very similar.
Additionally we estimate the imposed norm separately for the first ten and the remaining rounds in Appendix A.4. Again, results are robust.

	Imposed Norm	
Constant	11.33*** (0.95)	16.41*** (2.43)
HighTrans	−0.65 (1.35)	−0.43 (1.40)
HighPwr	0.75 (1.35)	1.01 (1.38)
HighTrans x HighPwr	−0.28 (1.91)	−0.61 (1.95)
t	0.18*** (0.01)	0.10*** (0.02)
t x HighTrans	−0.01 (0.01)	−0.004 (0.01)
t x HighPwr	0.03* (0.01)	0.04* (0.01)
t x HighTrans x HighPwr	−0.10*** (0.02)	−0.10*** (0.02)
Controls	×	✓
Observations	2,880	2,880
Log Likelihood	−6,727.82	−6,753.32
Akaike Inf. Crit.	13,475.65	13,550.64
Bayesian Inf. Crit.	13,535.30	13,681.88
Notes:	$^{\cdot}p < 0.1; ^{*}p < 0.05; ^{**}p < 0.01; ^{***}p < 0.001;$	

Table 3: Linear mixed-effects model of the imposed norm.
Controls include age, gender, spite, SVO, narcissism, rivalry, and the interaction of those variables with the period.

Hence, it seems like the abusive behavior is driven by the norm punishers imposed. The imposed norm is strongest in the high-power, low-transparency treatment, and the norm is the lowest in the high-power, high-transparency treatment. Under both transparency settings low power leads to an intermediate imposed norm which does not differ significantly between the transparency settings.

4.3. Contribution behavior

In a last step, the contribution behavior of all participants will be analyzed. Figure 3 illustrates a significant difference in contributions between punishers and non-punishers. Comparing the average contribution behavior over all rounds reveals that punishers contributed only M= 9.38 ($\overline{SD}^{14}$= 3.57) while non-punishers contributed M= 16.19 ($\overline{SD}$= 4.28) points to the public-good, a highly significant difference: $t(190)$= -7.9, $p \leq 0.01$.[15]

Discriminating by treatments in Figure 3, we observe that high power does not lead to lower contributions by the punisher, nor does increased transparency induce higher contributions per se. However, more transparency does improve contributions when combined with high power. The combination of both treatments is also the sole one to see an increase in contributions by punishers over time.

Non-punishers start with similar contributions in all treatments and increase their contributions over time. While high transparency generally strengthens the increase in contributions, the combination with high power dampens this positive effect significantly.[16]

Treatment effects are reported in Table 4 via a mixed-effects linear regression with fixed effects for the treatments and random effects for subjects and groups, while controlling for the contribution in the last round. In Table 3, we also included age, gender, spite, rivalry, narcissism and SVO, and the interaction of the mentioned factors with time as controls. All results are robust to those controls.[17]

Result 3 a Punishers contribute far less then non-punishers.

[14] Average of the standard deviations over all rounds.

[15] This effect has also been reported in Hoeft and Mill (2017) for the low-power treatments.

[16] Note that we do not necessarily argue that the non-punisher's behavior is directly driven by the treatments. It might well be that the change in contribution is due to differently imposed norms, which are directly influenced by the treatments. We nevertheless want to describe how contribution behavior changes, directly or indirectly, due to the treatments.

[17] We use a linear model as this is a common approach in the estimation of contribution behavior in public-goods experiments (see, for example, Fehr and Gächter, 2002, Nikiforakis, 2010, Nikiforakis and Normann, 2008). However, like in the two previous sections, we also relax this assumption in Appendix B.1.1 and Appendix B.1.2. For that purpose, we estimate a common loess contribution spline over rounds over all treatments. Using a Bayesian approach, we compare the curvity of the contribution over time of all the treatments. The results are qualitatively identical to the results reported here.

Result 3 b Punishers increase their contributions over time only in the high-power, high-transparency treatment.

Result 3 c Non-punishers increase their contributions over time. Increased transparency leads to a stronger increase overall.

Result 3 d High power dampens the positive effect of transparency on the contributions of non-punishers.

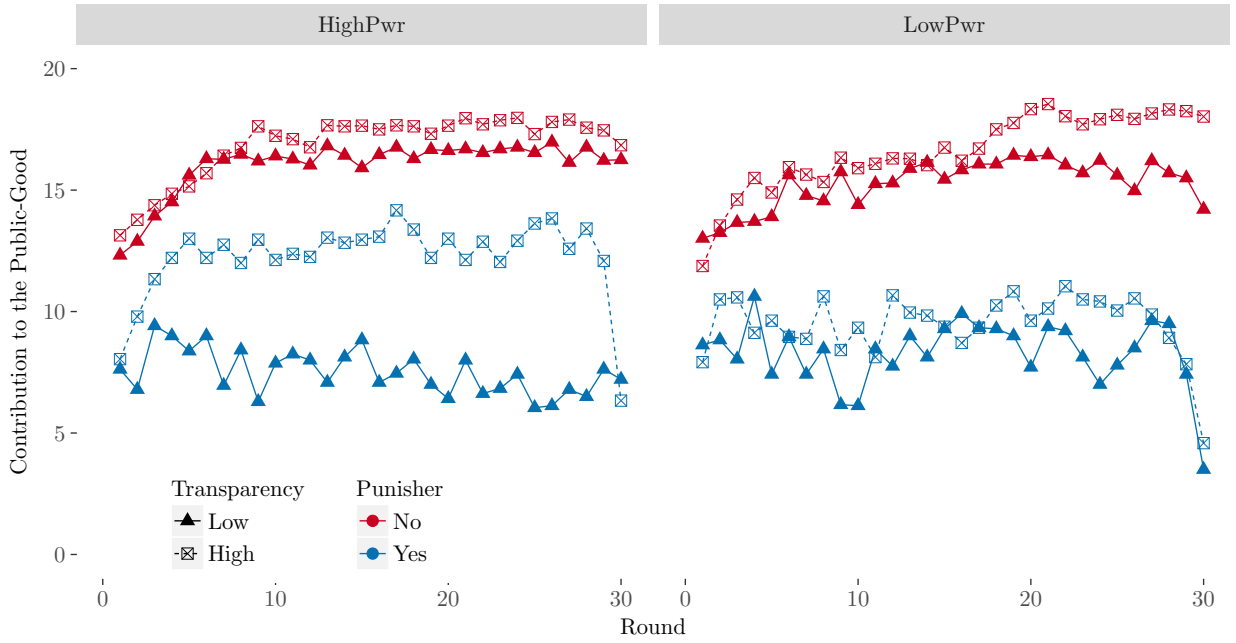

Figure 3: Contribution to the public-good over time in the respective treatments.
Blue lines represent the punisher's contributions, while red lines represent the non-punisher's contributions. The graph on the right shows the contribution behavior in the low-power treatments, while the graph on the left shows the high-power treatments. The high-transparency treatments are represented by dashed lines with crossed cubes, while the low-transparency treatments are shown with solid lines and solid triangles.

	Contribution to Public-Good			
	Non-Punisher		Punisher	
Constant	14.08*** (0.82)	13.63*** (1.22)	8.35*** (1.54)	5.36 (3.87)
HighTrans	−0.08 (1.17)	−0.18 (1.17)	1.10 (2.18)	1.27 (2.23)
HighPwr	0.55 (1.17)	0.48 (1.16)	−0.27 (2.18)	−1.14 (2.20)
HighTrans x HighPwr	0.46 (1.65)	0.46 (1.65)	2.34 (3.09)	2.76 (3.11)
t	0.08*** (0.01)	0.15*** (0.03)	0.005 (0.02)	−0.05 (0.05)
t x HighTrans	0.09*** (0.01)	0.09*** (0.01)	0.01 (0.03)	0.01 (0.03)
t x HighPwr	0.01 (0.01)	0.01 (0.01)	−0.03 (0.03)	−0.02 (0.03)
t x HighTrans x HighPwr	−0.06** (0.02)	−0.06** (0.02)	0.08* (0.04)	0.07· (0.04)
LastRound	−1.61*** (0.25)	−1.61*** (0.25)	−4.28*** (0.49)	−4.28*** (0.49)
Controls	×	✓	×	✓
Observations	8,640	8,640	2,880	2,880
Log Likelihood	−24,720.45	−24,734.21	−8,610.14	−8,629.57
Akaike Inf. Crit.	49,464.90	49,516.41	17,244.27	17,307.14
Bayesian Inf. Crit.	49,549.67	49,685.95	17,315.86	17,450.31
Notes:				·p < 0.1;*p < 0.05;**p < 0.01;***p < 0.001;

Table 4: Linear mixed-effects model of the contribution to the public-good.
Controls include age, gender, spite, SVO, narcissism, rivalry, and the interaction of those variables with the period.
Standard errors are in brackets.

Summarizing the contribution behavior, we can see that the contribution of punishers is increasing only in the high-power, high-transparency treatment, while in all other treatments the contribution stays virtually the same.

Hence, the treatment differences in abusive behavior are mainly driven by the imposed norm. However, the difference in contribution behavior strengthens the effect of transparency in the high power setting.

5. Discussion

Modern societies rely heavily on institutionalized punishment. The power conferred in these institutions can be abused in different ways: One is to harm others directly. A more pervasive and frequent phenomenon is to use the institutional power for selfish gains by undercutting the institutional purpose, thereby harming others indirectly. This is often done by enforcing rules that are not adhered to by the institutional members. Examples range from illegal violence by police to unjust enrichment or tax evasion by government elites. We provide a first investigation of institutional second-party punishers in a repeated game and indeed find frequent and strong abuse of power across a large majority of subjects.

In line with our hypothesis, we find that a (theoretically inconsequential) increase in punishment power leads to participants learning to abuse their power faster over time. The results of our transparency treatments, however, run counter to our hypothesis: Instead of limiting abuse for low-power treatments, it did so only for the high power one. This is an interesting finding that suggests the relationship between power and transparency is more complex than previously thought. We conjecture that this finding can be explained in two possible ways: On the one hand, transparency might make moral features of a situation more salient. While in the low-power treatment the punisher might not realize that he has enough power to deter everybody from free-riding, the high powered punisher will quickly realize that he is in a position to dominate others completely. This might trigger moral or empathy driven effects. Another, somewhat less plausible, explanation would be that high powered punishers realize their privileged standing under transparency and fear a revolt due to the lack of fairness. Previous experiments show that participants are especially sensitive to discrepancies of power and tend to ostracize those who wield it if possible (Ramalingam et al., 2015). There may be evolutionary reasons why humans strongly reject stark dominance hierarchies that are not culturally justified (Gintis, 2016).

This should caution against the - at times overly optimistic - picture painted by the vast literature on prosocial punishment. Not only a small subset, but a large part of the population were willing to bypass their own norm. Further research should improve our knowledge on the complex relationship of (power) abuse and transparency and investigate under what circumstances resistance is organized.

Acknowledgements

We thank Urs Fischbacher, Nikos Nikiforakis, Martin Kocher, Ulrike Malmendier, Alexander Vostroknutov, Eugenio Verrina, Julia Sasse, Daniel Houser, Werner Güth, Oliver Kirchkamp, and Christoph Engel for helpful comments. We appreciate comments from the participants of the Economic Science Association World Conference 2016, the Jena Econ-Seminar and the Bonn Econ-Seminar. We gratefully acknowledge funding from the Max Planck Society and the IMPRS-Uncertainty.

References

Ambrus, A., Greiner, B., 2012. Imperfect public monitoring with costly punishment: an experimental study. American Economic Review 102 (7), 3317–32.

Andreoni, J., Croson, R., 2008. Partners versus strangers: Random rematching in public goods experiments. In: Plott, C. R., Smith, V. L. (Eds.), Handbook of Experimental Economics Results, 1st Edition. Vol. 1, Part 6. Elsevier, Ch. 82, pp. 776–783.

Andreoni, J., Gee, L., 2012. Gun for hire: Delegated enforcement and peer punishment in public goods provision. Journal of Public Economics 96 (11), 1036–1046.

Azrieli, Y., Chambers, C. P., Healy., P. J., 2015. Incentives in experiments: A theoretical analysis, mimeo.

Back, M. D., Kuefner, A. C. P., Dufner, M., Gerlach, T. M., Rauthmann, J. F., Denissen, J. J. A., 2013. Narcissistic admiration and rivalry: Disentangling the bright and dark sides of narcissism. Journal of Personality and Social Psychology 105 (10), 1013–1037.

Baldassarri, D., Grossman, G., 2011a. Centralized sanctioning and legitimate authority promote cooperation in humans. Proceedings of the National Academy of Sciences 108 (27), 11023–11027.

Baldassarri, D., Grossman, G., 2011b. Centralized sanctioning and legitimate authority promote cooperation in humans. Proceedings of the National Academy of Sciences 108, 11023–11027.

Bock, O., Baetge, I., Nicklisch, A., 2014. hroot: Hamburg registration and organization online tool. European Economic Review 71 (C), 117–120.

Carpenter, J., Matthews, P. H., 2009. What norms trigger punishment? Experimental Economics 12 (3), 272–288.

Carpenter, J. P., Matthews, P. H., 2012. Norm enforcement: anger, indignation, or reciprocity? Journal of the European Economic Association 10 (3), 555–572.

Cinyabuguma, M., Page, T., Putterman, L., 2006. Can second-order punishment deter perverse punishment? Experimental Economics 9 (3), 265–279.

Croson, R. T., 2001. Feedback in voluntary contribution mechanisms: An experiment in team production. In: Isaac, R. M., Norton, D. A. (Eds.), Research in Experimental Economicss. Vol. 8 of Research in Experimental Economics. Emerald Group Publishing Limited, pp. 85–97.

Engel, C., Zhurakhovska, L., 2017. You are in charge: Experimentally testing the motivating power of holding a judicial office. The Journal of Legal Studies 46 (1), 1–50.

Falk, A., Fehr, E., Fischbacher, U., 2005. Driving forces behind informal sanctions. Econometrica 73 (6), 2017–2030.

Fehr, E., Fischbacher, U., 2004. Third-party punishment and social norms. Evolution and Human Behavior 25 (2), 63–87.

Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415 (6868), 137–140.

Fehr, E., Williams, T., 2013. Endogenous emergence of institutions to sustain cooperation. Tech. rep., Working Paper.

Fischbacher, U., 2007. z-tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10 (2), 171–178.

Fischer, S., Grechenig, K., Meier, N., 2016. Monopolizing sanctioning power under noise eliminates perverse punishment but does not increase cooperation. Frontiers in Behavioral Neuroscience 10.

Gintis, H., 2016. Individuality and entanglement: the moral and material bases of social life. Princeton University Press.

Gross, J., Mder, Z. Z., Okamoto-Barth, S., Riedl, A., 2016. Building the leviathan - voluntary centralisation of punishment power sustains cooperation in humans. Scientific Reports 6, 20767.

Guala, F., 2012. Reciprocity: Weak or strong? what punishment experiments do demonstrate. Behavioral and Brain Sciences 35 (1), 1–15.

Herrmann, B., Thöni, C., Gächter, S., 2008. Antisocial punishment across societies. Science 319 (5868), 1362–1367.

Hilbe, C., Traulsen, A., Röhl, T., Milinski, M., 2013. Democratic decisions establish stable authorities that overcome the paradox of second-order punishment. Proceedings of the National Academy of Sciences 111 (2), 201315273–756.

Hoeft, L., Mill, W., 2017. Selfish punishers. Economic Letters 157, 41–44.

Houser, D., Xiao, E., 2010. Inequality-seeking punishment. Economics Letters 109 (1), 20–23.

Johnson, T., Dawes, C. T., Fowler, J. H., McElreath, R., Smirnov, O., 2009. The role of egalitarian motives in altruistic punishment. Economics Letters 102 (3), 192–194.

Khadjavi, M., Lange, A., Nicklisch, A., 2014. The social value of transparency and accountability: experimental evidence from asymmetric public goods games. WiSo-HH Working Paper 1 (12).

Leibbrandt, A., López-Pérez, R., 2012. An exploration of third and second party punishment in ten simple games. Journal of Economic Behavior & Organization 84 (3), 753–766.

Marcus, D. K., Zeigler-Hill, V., Mercer, S. H., Norris, A. L., 2014. The psychology of spite and the measurement of spitefulness. Psychological Assessment 26 (2), 563–574.

Markussen, T., Putterman, L., Tyran, J.-R., 2014. Self-organization for collective action: An experimental study of voting on sanction regimes. The Review of Economic Studies 81 (1), 301–324.

Murphy, R. O., Ackerman, K. A., 2014. Social value orientation: Theoretical and measurement issues in the study of social preferences. Personality and Social Psychology Review 18 (1), 13–41.

Murphy, R. O., Ackerman, K. A., Handgraaf, M. J. J., 2011. Measuring social value orientation. Judgment and Decision Making 6 (8), 771–781.

Nicklisch, A., Grechenig, K., Thöni, C., 2016. Information-sensitive Leviathans. Journal of Public Economics 144, 1–13.

Nikiforakis, N., 2010. Feedback, punishment and cooperation in public good experiments. Games and Economic Behavior 68 (2), 689–702.

Nikiforakis, N., Normann, H.-T., 2008. A comparative statics analysis of punishment in public-good experiments. Experimental Economics 11 (4), 358–369.

O'Gorman, R., Henrich, J., Van Vugt, M., 2009. Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. Proceedings of the Royal Society of London B: Biological Sciences 276 (1655), 323–329.

Patel, A., Cartwright, E., van Vugt, M., 2010. Punishment cannot sustain cooperation in a public good game with free-rider anonymity, mimeo.

Putterman, L., Tyran, J.-R., Kamei, K., 2011. Public goods and voting on formal sanction schemes. Journal of Public Economics 95 (910), 1213–1222.

Ramalingam, A., Rojo Arjona, D., Schram, A., Van Leeuwen, B., 2015. Authority and Centrality: Power and Cooperation in Social Dilemma Networks. IAST Working Papers 15-23, Institute for Advanced Study in Toulouse (IAST).

Reuben, E., Riedl, A., 2013. Enforcement of contribution norms in public good games with heterogeneous populations. Games and Economic Behavior 77 (1), 122–137.

Traulsen, A., Röhl, T., Milinski, M., 2012. An economic experiment reveals that humans prefer pool punishment to maintain the commons. Proceedings of the Royal Society of London B: Biological Sciences 279 (1743), 3716–3721.

Xiao, E., 2013. Profit-seeking punishment corrupts norm obedience. Games and Economic Behavior 77 (1), 321–344.

Xiao, E., Tan, F., 2014. Justification and legitimate punishment. Journal of Institutional and Theoretical Economics 170 (1), 168–188.

Zhang, B., Li, C., Silva, H., Bednarik, P., Sigmund, K., 2014. The evolution of sanctioning institutions: an experimental approach to the social contract. Experimental Economics 17 (2), 285–303.

Zhou, Y., Jiao, P., Zhang, Q., 2017. Second-party and third-party punishment in a public goods experiment. Applied Economics Letters 24 (1), 54–57.

Appendix A. Additional regressions

Appendix A.1. Alternative definition of abuse

In the paper, we stipulate that the abuse of the punishment power is described by the deviation of the punisher's contribution from the norm he imposes upon non-punishers.

However, one might argue that abusive behavior could already be described by the deviation of the punisher's contribution from the average contribution of non-punishers.

It is worth pointing out several shortcomings of this approach.

- This alternative approach does not take punishment into account. Hence, subjects are considered abusive even if they have never forced any other group member to contribute anything to the public-good (in case they contribute less then the average non-punisher). Similarly, subjects who do punish the other group members are not considered abusive if the non-punishers are not sensitive to punishment and contribute more than or equal to the punisher. In these instances, the average group contribution might be driven not so much by the behavior of the punisher, but by social preferences etc.

- It is problematic to compare the contribution of the punisher to the average contribution of the non-punishers as the punisher does not know how much the non-punishers will contribute in any given round. The punisher can at most adjust his contribution to the average contribution of the other from the previous rounds.

Given those drawbacks, we advocate the definition of abusive behavior as the deviation of the punisher's contribution from the norm he imposes.

Nevertheless, we use this alternative approach to estimate this deviation with a linear[18] mixed-effects model similar to section 4.1. The results can be found in Table A.5. The results are very similar to the results from section 4.1. The only qualitative difference is that under low transparency high power does not lead to stronger abuse over time. Otherwise, the remaining results prevail.

	Abuse	
Constant	5.73*** (1.48)	8.46* (3.79)
HighTrans	−1.18 (2.09)	−1.43 (2.18)
HighPwr	0.82 (2.09)	1.30 (2.16)
HighTrans x HighPwr	−1.88 (2.95)	−2.19 (3.05)
t	0.08*** (0.02)	0.08 (0.05)
t x HighTrans	0.09** (0.03)	0.09** (0.03)
t x HighPwr	0.05 (0.03)	0.04 (0.03)
t x HighTrans x HighPwr	−0.14*** (0.04)	−0.14** (0.04)
LastRound	2.67*** (0.53)	2.67*** (0.53)
Controls	×	✓
Observations	2,880	2,880
Log Likelihood	−8,851.44	−8,872.72
Akaike Inf. Crit.	17,724.88	17,791.43
Bayesian Inf. Crit.	17,790.50	17,928.64
Notes:	$^{\cdot}$p < 0.1;*p < 0.05;**p < 0.01;***p < 0.001;	

Table A.5: Mixed-effects model of the Amount of Abuse in the alternative definition of abuse (the deviation of the punisher's contribution from the average contribution of the non-punishers).
Controls include age, gender, spite, SVO, narcissism, rivalry, and the interaction of those variables with the period.

Appendix A.2. Punishment behavior

In this section, we examine the punishment behavior. In estimating the punishment behavior we follow mainly the standard approach commonly used in public-goods experiments with punishment (e.g., Fehr and Gächter, 2002, Nikiforakis, 2010, Nikiforakis and Normann, 2008) - namely, estimating the punishment amount as a function of the contribution of the punisher and the contribution of the punished member. Note that punishers were not able to punish themselves, and hence we have three decisions per round for every punisher.

To estimate the punishment given to non-punishers we compare the contribution of those subjects to the average of non-punishers. Note that mostly in the literature the contribution of the punishment receiving subjects is compared either to the individual contribution of the punisher[19] or to the average contribution of the whole group.[20] However, comparing model fits of these two alternative approaches to our implemented approach shows unequivocally that our approach dominates. The log-likelihood of the model comparing the contribution of the punished subject to average contribution of non-punisher is -2.4423×10^4 and is significantly better than the log-likelihood of the model comparing the contribution of the punished subject to average contribution of all subjects with -2.4799×10^4 ($\chi^2(0)= 1507.21$, $p \leq 0.01$), and also compared to a model which compares the contribution of the punished subject to the punishing person (loglik: -2.7208×10^4) ($\chi^2(0)= 11156.418$, $p \leq 0.01$).

[18] Similarly to previous sections, we also estimate this part with a loess spline in a Bayesian framework in section Appendix B.3.2.

[19] For example, see Fehr and Gächter (2002) or Reuben and Riedl (2013).

[20] For example, see Carpenter and Matthews (2009, 2012).

We use the following econometric model to estimate punishment behavior:

$$
\begin{aligned}
Pun_{i,k,t} =& \beta_0 + \beta_1 \cdot \text{HighPwr} + \beta_2 \cdot \text{HighTrans} + \beta_3 \cdot \text{HighPwr} \cdot \text{HighTrans} \\
& \beta_4 \cdot t + \beta_5 \cdot t \cdot \text{HighPwr} + \beta_6 \cdot t \cdot \text{HighTrans} + \beta_7 \cdot t \cdot \text{HighPwr} \cdot \text{HighTrans} \\
& \beta_8 \cdot (\overline{c_{i\neq D,k,t}} - c_{i,k,t})^+ + \beta_9 \cdot \text{HighPwr} \cdot (\overline{c_{i\neq D,k,t}} - c_{i,k,t})^+ + \\
& \beta_{10} \cdot \text{HighTrans} \cdot (\overline{c_{i\neq D,k,t}} - c_{i,k,t})^+ + \beta_{11} \cdot \text{HighPwr} \cdot \text{HighTrans} \cdot (\overline{c_{i\neq D,k,t}} - c_{i,k,t})^+ \\
& \beta_{12} \cdot t \cdot (\overline{c_{i\neq D,k,t}} - c_{i,k,t})^+ + \beta_{13} \cdot t \cdot \text{HighPwr} \cdot (\overline{c_{i\neq D,k,t}} - c_{i,k,t})^+ + \\
& \beta_{14} \cdot t \cdot \text{HighTrans} \cdot (\overline{c_{i\neq D,k,t}} - c_{i,k,t})^+ + \beta_{15} \cdot t \cdot \text{HighPwr} \cdot \text{HighTrans} \cdot (\overline{c_{i\neq D,k,t}} - c_{i,k,t})^+ \\
& + \epsilon_i + \epsilon_{i,k} + \epsilon_{i,k,t}
\end{aligned} \tag{A.1}
$$

where HighPwr is a dummy with value one for the high-power treatment and zero otherwise. HighTrans is a dummy with value one if the subject was in the high-transparency treatment and zero otherwise. $\overline{c_{i\neq D,k,t}}$ denotes the average contribution of all non-punishers in group k in round t and $(\overline{c_{i\neq D,k,t}} - c_{i,k,t})^+$ denotes the maximum of 0 and the deviation of the contribution of subject i from the average contribution of the non-punishers. ϵ_i and $\epsilon_{i,k}$ are the level 1 and level 2 random effects of groups and individuals.

The results of the estimation are presented in Table A.6. It can be seen that the deviation from the average contribution of the other non-punishers is punished and that this deviation is punished significantly stronger under high power and also under high transparency. However, the effect of transparency is stronger under low power then under high power.

Overall punishment was reduced over time even though punishment strength increased. Under low power, transparency led to a smaller increase in punishment strength over time then under high power.

Result 4 a The deviation from the average contribution of non-punishers was punished harshly and increased over time.

Result 4 b Power as well as transparency lead to harsher punishment for the deviation from the average contribution of other non-punishers.

Result 4 c The effect of transparency is stronger under low power than under high power.

Result 4 d Under low power, high transparency leads to a less severe increase in punishment over time.

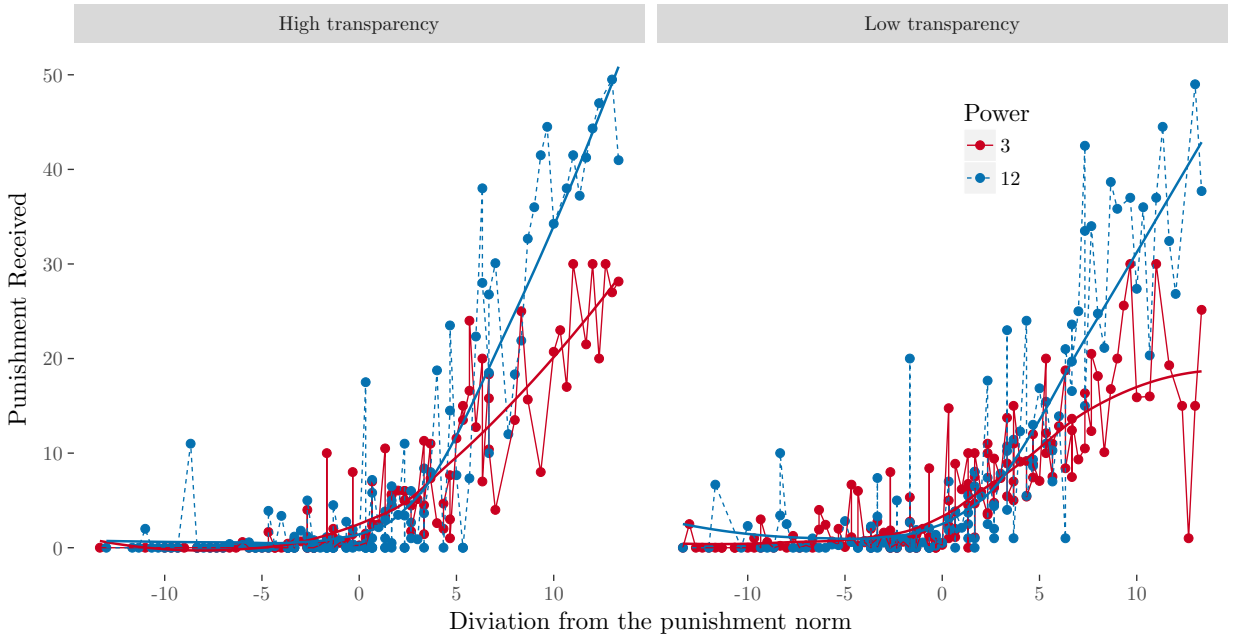

Figure A.4: Punishment received for deviating from the average contribution of the non-punishers with loess splines. Blue lines represent the received punishment in the high-power treatments, while red lines denote the low-power treatments. The graph on the right shows the received punishment in the low-transparency treatments and the high-transparency treatments are show on the right.

Constant	1.58*** (0.47)
HighPwr	−0.06 (0.66)
HighTrans	−0.77 (0.66)
HighPwr x HighTrans	0.18 (0.93)
$\overline{c_N} - c_j$	1.57*** (0.06)
$\overline{c_N} - c_j$ x HighPwr	1.11*** (0.09)
$\overline{c_N} - c_j$ x HighTrans	0.54*** (0.09)
$\overline{c_N} - c_j$ x HighPwr x HighTrans	−0.33* (0.13)
t	−0.05*** (0.01)
t x HighPwr	0.003 (0.02)
t x HighTrans	0.02 (0.02)
t x HighPwr x HighTrans	−0.02 (0.02)
t x $\overline{c_N} - c_j$	0.01*** (0.003)
t x $\overline{c_N} - c_j$ x HighPwr	−0.005 (0.01)
t x $\overline{c_N} - c_j$ x HighTrans	−0.02*** (0.01)
t x $\overline{c_N} - c_j$ x HighPwr x HighTrans	0.02* (0.01)
Observations	8,640
Log Likelihood	−24,423.49
Akaike Inf. Crit.	48,884.97
Bayesian Inf. Crit.	49,019.19
Notes:	·p<0.1;*p<0.05;**p<0.01;***p < 0.001;

Table A.6: Punishment behavior of punishers with respect to the four treatments and the punished subjectes' contribution deviation from the average contribution of all non-punishers.

Appendix A.3. Punishment Norms

This section extends Appendix A.2 to get a more detailed picture of punishment by estimating punishment norms similar to Carpenter and Matthews (2009), Reuben and Riedl (2013). To get a better picture, we estimate for each treatment which punishment norm was enforced. Note that we estimate a treatment

average. The enforced punishment norm is therefore different from the imposed contribution in that we estimate punishment for treatments, while we estimate the imposed contribution norm per subject. Hence, the imposed contribution is more efficient. Since the definition of abuse refers to the misuse of an entity for personal gain, it was necessary to investigate individual decision-making. Here we are interested in the aggregate effects of our treatments, understood as institutions, and extend Appendix A.2 by estimating the punishment norm and additionally modeling punishment as a two-step-process. For that purpose we differentiate between the decision to punish at all and the decision on how much to punish conditional on the decision to punish. Hence, we assume that subjects first make a decision to punish or not to punish. They punish with probability $1-\omega$, and with probability ω subjects are not punished. If subjects are punished, then the punishment amount follows a normal distribution.[21]

As subjects make two decisions, we allow two separate punishment norms, namely the punishment norm on whether to punish (called γ^v) and the punishment norm on how much to punish (called γ^p).

We use the following econometric model similar to Carpenter and Matthews (2009):

$$\begin{aligned} p^*_{i,k,t} =& \beta_0 + \beta_1 \cdot \text{HighPwr} + \beta_2 \cdot \text{HighTrans} + \beta_3 \cdot \text{HighPwr} \cdot \text{HighTrans} \\ & \beta_4 \cdot t + \beta_5 \cdot t \cdot \text{HighPwr} + \beta_6 \cdot t \cdot \text{HighTrans} + \beta_7 \cdot t \cdot \text{HighPwr} \cdot \text{HighTrans} \\ & \beta_8 \cdot (\gamma^p - c_{i,k,t})^+ + \beta_9 \cdot \text{HighPwr} \cdot (\gamma^p - c_{i,k,t})^+ + \\ & \beta_{10} \cdot \text{HighTrans} \cdot (\gamma^p - c_{i,k,t})^+ + \beta_{11} \cdot \text{HighPwr} \cdot \text{HighTrans} \cdot (\gamma^p - c_{i,k,t})^+ \\ & \beta_{12} \cdot t \cdot (\gamma^p - c_{i,k,t})^+ + \beta_{13} \cdot t \cdot \text{HighPwr} \cdot (\gamma^p - c_{i,k,t})^+ + \\ & \beta_{14} \cdot t \cdot \text{HighTrans} \cdot (\gamma^p - c_{i,k,t})^+ + \beta_{15} \cdot t \cdot \text{HighPwr} \cdot \text{HighTrans} \cdot (\gamma^p - c_{i,k,t})^+ \\ & + \epsilon_i + \epsilon_{i,k} + \epsilon_{i,k,t} \end{aligned} \tag{A.2}$$

$$\begin{aligned} v^*_{i,k,t} = g(\omega^*_{i,k,t}) =& \alpha_0 + \alpha_1 \cdot \text{HighPwr} + \alpha_2 \cdot \text{HighTrans} + \alpha_3 \cdot \text{HighPwr} \cdot \text{HighTrans} \\ & \alpha_4 \cdot t + \alpha_5 \cdot t \cdot \text{HighPwr} + \alpha_6 \cdot t \cdot \text{HighTrans} + \alpha_7 \cdot t \cdot \text{HighPwr} \cdot \text{HighTrans} \\ & \alpha_8 \cdot (\gamma^v - c_{i,k,t})^+ + \alpha_9 \cdot \text{HighPwr} \cdot (\gamma^v - c_{i,k,t})^+ + \\ & \alpha_{10} \cdot \text{HighTrans} \cdot (\gamma^v - c_{i,k,t})^+ + \alpha_{11} \cdot \text{HighPwr} \cdot \text{HighTrans} \cdot (\gamma^v - c_{i,k,t})^+ \\ & \alpha_{12} \cdot t \cdot (\gamma^v - c_{i,k,t})^+ + \alpha_{13} \cdot t \cdot \text{HighPwr} \cdot (\gamma^v - c_{i,k,t})^+ + \\ & \alpha_{14} \cdot t \cdot \text{HighTrans} \cdot (\gamma^v - c_{i,k,t})^+ + \alpha_{15} \cdot t \cdot \text{HighPwr} \cdot \text{HighTrans} \cdot (\gamma^v - c_{i,k,t})^+ \\ & + \epsilon_i + \epsilon_{i,k} + \epsilon_{i,k,t} \end{aligned} \tag{A.3}$$

$$\begin{aligned} & v_{i,k,t} = 1 \text{ if } v^*_{i,k,t} > 0 \\ & p_{i,k,t} = p^*_{i,k,t} \cdot v_{i,k,t} \end{aligned}$$

where $(x)^+ = max(x, 0)$. The norm to enforce is γ^v. The norm on how much to punish is γ^p. HighPwr is a dummy with value one if the treatment is the high-power treatment, and zero otherwise. HighTrans is a dummy with value one if the treatment is the high-transparency treatment, and zero otherwise. t is the period. $v_{i,k,t}$ is a variable indicating whether subject i in group k was punished in period t. $p_{i,k,t}$ is a variable indicating how much subject i in group k was punished in period t. $c_{i,k,t}$ denotes how much subject i contributed in period t to the group account of group k. ϵ_i and $\epsilon_{i,k}$ are the level 1 and level 2 random effects of groups and individuals. $g(\cdot)\colon (0,1) \to \mathbb{R}$ is the link function. We will use the logit link: $g(\mu) = log(\mu/(1-\mu))$, as this is easier to interpret.

[21] Note: As the punishment power is obviously bigger in the high-power treatment, we will look at the punishment amount relative to the maximally possible amount. Otherwise, we might find differences between the power treatments basically by definition.

To sum up: we estimate how the deviation of a subjects' contribution from the punishment norm influences the decision to punish at all and, if so, how severely.

The possible punishment norms we took into account were: the average contribution of the whole group, the average contribution of non-punishers, the contribution of the punisher, and absolute norms ($\in \{0, 0.1, ..., 19.8, 19.9, 20\}$). Table A.7 shows the log likelihoods of the estimations with the possible norms in each treatment.

Several points are worth mentioning. First, it is evident that the best absolute norm for the decision whether to punish is 20 in each treatment. Since many punishers undercut this norm, it suggests abuse on the aggregate level. This norm also performs best for the decision to punish in all treatments but the high-power, low-transparency treatment (where the average contribution of non-punishers is the best norm).

It is also interesting to see that the contribution of the punisher is the worst norm in any treatment in either decision, which in combination with the much lower average contribution of punishers is also indicative of abuse.[22] Also, the absolute norm does not perform best in the decision how much to punish.

It is noteworthy that the norm on how much to punish is the average contribution of non-punishers in all treatments but the high-power, high-transparency treatment. This treatment is the only treatment which takes the average contribution of the whole group as the norm, which is remarkable, as this implies that the average punisher in this treatment takes his own contribution also into account. This, again, supports the conclusion that under high power and high transparency, punishers behave less abusively (as they do not differentiate between themselves and the non-punishers).

Table A.8 reports the estimates of equation A.2 and A.3 for each treatment separately with the relative best norms. Table A.9 reports the estimates of equation A.2 and A.3 with the relative best norms.

	LowPwr ∧ LowTrans		LowPwr ∧ HighTrans		HighPwr ∧ LowTrans		HighPwr ∧ HighTrans	
	Punish?	Punishment	Punish?	Punishment	Punish?	Punishment	Punish?	Punishment
ContributionPun	-828	-10	-696	-4	-832	250	-451	112
ContAvgNonPun	-542	112*	-539	131*	-588*	344*	-403	180
ContAvg	-642	79	-582	101	-665	343	-421	188*
Best Absolut Norm	-541*	55	-501*	57	-627	307	-296*	163
Best Absolut Norm Value	20	11.3	20	19.1	20	0.1	20	12.4

Note: Log likehoods of the equation A.3 and A.2 with the specificed punishment norms

Table A.7: Loglik of the norms suggested for punishment. The best norm is expressed by *.

[22] Note that Reuben and Riedl (2013) base their investigation on exactly the deviation of the punished subjects' contribution from the punisher's contribution.

	Punishment Behavior							
	Punish?				Punishment amount			
	HighPwr HighTrans	LowPwr HighTrans	HighPwr LowTrans	LowPwr LowTrans	HighPwr HighTrans	LowPwr HighTrans	HighPwr LowTrans	LowPwr LowTrans
Constant	−6.34***	−3.55***	−1.54***	−2.83***	0.03	0.15***	0.07*	0.22***
	(1.07)	(0.60)	(0.35)	(0.48)	(0.03)	(0.04)	(0.03)	(0.05)
$\left(\gamma^* - c_{i,k,t}\right)^+$	0.90***	0.45***	0.93***	0.42***	0.03***	0.05***	0.02***	0.03***
	(0.08)	(0.04)	(0.11)	(0.03)	(0.003)	(0.004)	(0.003)	(0.004)
t	−0.06**	−0.11***	−0.07***	−0.07***	−0.001	0.001	−0.0002	0.002
	(0.02)	(0.02)	(0.01)	(0.01)	(0.001)	(0.001)	(0.001)	(0.001)
t x $\left(\gamma^* - c_{i,k,t}\right)^+$	0.01**	0.02***	0.01·	−0.0004	0.001***	−0.0003	0.001***	0.001*
	(0.003)	(0.003)	(0.01)	(0.002)	(0.0002)	(0.0003)	(0.0001)	(0.0002)
Model	Logistic Mixed Effects	Logistic Mixed Effects	Logistic Mixed Effects	Logistic Mixed Effects	Linear Mixed Effects	Linear Mixed Effects	Linear Mixed Effects	Linear Mixed Effects
Cond. on pun.	×	×	×	×	✓	✓	✓	✓
Observations	2,160	2,160	2,160	2,160	326	386	556	549
Log Likelihood	−296.28	−500.67	−587.74	−541.31	187.52	131.14	344.15	111.81
Akaike Inf. Crit.	602.55	1,011.35	1,185.49	1,092.62	−363.04	−250.28	−676.30	−211.62
Bayesian Inf. Crit.	630.94	1,039.73	1,213.88	1,121.01	−340.32	−226.54	−650.37	−185.77
Note:								·p < 0.1;*p < 0.05;**p < 0.01;***p < 0.001;

Table A.8: Mixed-effects estimates for punishment following the individually best norms for the respective treatments as reported in Table A.7.

	Punishment Behavior	
	Punish?	Punishment amount
Constant	−2.75*** (0.57)	0.22*** (0.04)
HighPwr	1.15 (0.79)	−0.15** (0.05)
HighTrans	−0.73 (0.83)	−0.07 (0.05)
HighPwr x HighTrans	−2.75* (1.20)	0.04 (0.07)
$\overline{c_N} - c_j$	0.40*** (0.04)	0.03*** (0.003)
$\overline{c_N} - c_j$ x HighPwr	0.54*** (0.12)	−0.01** (0.004)
$\overline{c_N} - c_j$ x HighTrans	0.04 (0.05)	0.02*** (0.01)
$\overline{c_N} - c_j$ x HighPwr x HighTrans	−0.26· (0.14)	−0.01* (0.01)
t	−0.08*** (0.02)	0.002* (0.001)
t x HighPwr	−0.01 (0.02)	−0.002 (0.002)
t x HighTrans	−0.03 (0.02)	−0.002 (0.002)
t x HighPwr x HighTrans	0.03 (0.03)	0.001 (0.002)
t x $\overline{c_N} - c_j$	0.0001 (0.002)	0.0004* (0.0002)
t x $\overline{c_N} - c_j$ x HighPwr	0.02* (0.01)	0.0000 (0.0003)
t x $\overline{c_N} - c_j$ x HighTrans	0.02*** (0.003)	−0.001* (0.0003)
t x $\overline{c_N} - c_j$ x HighPwr x HighTrans	−0.03** (0.01)	0.001· (0.0004)
Model	Logistic Mixed Effects	Linear Mixed Effects
Cond. on pun.	×	✓
Observations	8,640	1,817
Log Likelihood	−1,917.78	740.03
Akaike Inf. Crit.	3,871.56	−1,442.05
Bayesian Inf. Crit.	3,998.72	−1,337.46
Note:		·p<0.1;*p<0.05;**p<0.01;***p < 0.001;

Table A.9: Mixed-effects estimates for punishment following the individual punishments norms for each treatment reported in Table A.7 while comparing the treatment effects.

Result 5 a Similarly to the findings by Carpenter and Matthews (2009): subjects use two distinct norms on the decision to punish and on the decision how much to punish.

Result 5 b For the decision to punish, the absolute norm of 20 (100% contribution) is performing best in all treatments but the low-transparency, high-power treatment, where the average contribution of non-punishers is the best-performing norm.

Result 5 c For the decision on how much to punish, the average contribution of non-punishers is used as the relative norm in all treatments but in the high power, high-transparency treatment. In this treatment the average contribution of the whole group (including the punisher) is the best-performing norm.

Result 5 d In all treatments, the contribution of the punishers is the worst-performing norm.

Appendix A.4. Imposed Norm

Similar to the main part of the paper, we regress the imposed norm for each of the four treatments. However, in contrast to the main part of the paper, we estimate two linear splines, with the first spline in the first ten periods and the second spline for the remaining periods.

It can be seen from Table A.10 that the results from the main part of the paper are driven mainly by the development in the first ten periods. The effects of the remaining periods show the same trends as in the first ten periods; however, these results are not robust to controls.

Hence, it can be concluded that the relevant norm was imposed within the first ten rounds and kept more or less constant in the remaining rounds. The same goes for the treatment differences, as all differences (as reported in the main part of the paper, see section 4.2) are developed mainly in the first ten rounds and kept constant afterwards.

	Imposed Norm			
	$t \in \{1, ..., 10\}$	$t \in \{11, ..., 30\}$	$t \in \{1, ..., 10\}$	$t \in \{11, ..., 30\}$
Constant	9.16^{***} (1.06)	13.65^{***} (0.98)	14.60^{***} (2.71)	19.39^{***} (2.49)
HighTrans	-0.77 (1.51)	-1.11 (1.38)	-0.91 (1.56)	-0.41 (1.43)
HighPwr	-0.11 (1.51)	0.93 (1.38)	-0.002 (1.54)	1.42 (1.42)
HighTrans x HighPwr	1.16 (2.13)	-1.41 (1.96)	1.11 (2.18)	-2.06 (2.00)
t	0.52^{***} (0.05)	0.07^{***} (0.01)	0.35^{**} (0.13)	-0.03 (0.03)
t x HighTrans	0.04 (0.08)	0.01 (0.01)	0.09 (0.08)	-0.01 (0.01)
t x HighPwr	0.20^{**} (0.08)	$0.02^{\cdot}$ (0.01)	0.22^{**} (0.08)	0.02 (0.01)
t x HighTrans x HighPwr	-0.34^{**} (0.11)	-0.05^{*} (0.02)	-0.39^{***} (0.11)	-0.03 (0.02)
Controls	×	×	✓	✓
Observations	960	1,920	960	1,920
Log Likelihood	−2,389.20	−3,550.38	−2,396.41	−3,535.96
Akaike Inf. Crit.	4,798.41	7,120.75	4,836.83	7,115.93
Bayesian Inf. Crit.	4,847.08	7,176.35	4,943.90	7,238.25
Notes:				$^{\cdot}$p < 0.1;*p < 0.05;**p < 0.01;***p < 0.001;

Table A.10: Linear mixed-effects model of the imposed norm split into first 10 periods and remaining periods. Controls include: Gender, Age, SVO, rivalry, narcissism, spite, and the interaction of those variables with the period.

Appendix B. Bayesian Approaches

In the main part of the paper, we used mainly linear regression, as this is a common approach. However, as can be seen, for example, in Figure 3, the contribution decision over time does not necessary evolve linearly. To relax this assumption, we will estimate loess splines over time to allow for non-linear behavior. We will take this approach for several dependent variables. For that purpose, we will estimate an average

loess spline for the respective dependent variables over time (average for all treatments)[23], which we call $\lambda(t)$.

The following general econometric model of a dependent variable Ind will be used in the rest of the paper:

$$\begin{aligned} Ind_{i,t} =& \beta_0 + \lambda(t) \cdot (1 + \beta_1 \cdot \mathbb{1}_{\text{HighPwr} \wedge \text{LowTrans}} + \beta_2 \cdot \mathbb{1}_{\text{HighPwr} \wedge \text{HighTrans}} \\ & + \beta_3 \cdot \mathbb{1}_{\text{LowPwr} \wedge \text{LowTrans}} + \beta_4 \cdot \mathbb{1}_{\text{LowPwr} \wedge \text{HighTrans}}) \\ & + \epsilon_i + \epsilon_{i,k,t} \end{aligned} \tag{B.1}$$

$Ind_{i,t}$ represents the dependent variable (contribution, abuse, etc) of subjects i in round t with $i \in \{1, \ldots, n\}$ and $t \in \{0, \ldots, 30\}$.

$\mathbb{1}_{\text{HighPwr} \wedge \text{LowTrans}}$ denotes a dummy variable which is one if the subject was in the high-power, low-transparency treatment, and zero otherwise. $\mathbb{1}_{\text{HighPwr} \wedge \text{HighTrans}}, \mathbb{1}_{\text{LowPwr} \wedge \text{LowTrans}}, \mathbb{1}_{\text{LowPwr} \wedge \text{HighTrans}}$ are defined accordingly.

To account for the nested structure of the data we included ϵ_i as the random effects of the individual i. $\epsilon_{i,k,t}$ is the residuals.

As vague priors we used:

$$\begin{gathered} \beta_{0,1,2,3,4} \sim \mathcal{N}(0, .0001) \\ \epsilon_i \sim \mathcal{N}(0, \tau_2)\,,\ \epsilon_{i,k,t} \sim \mathcal{N}(0, \tau_1) \tau_l \sim \Gamma(\frac{m_l^2}{s_l^2}, \frac{m_l^2}{s_l^2}) \\ \text{with } m_l \sim \Gamma(1,1)\,,\ s_l \sim \Gamma(1,1) \text{ with } l \in \{1,2\} \end{gathered}$$

The interpretation of the βs is the following: β_0 is the intercept. β_l with $l \in \{1, 2, 3, 4\}$ are the deviations in the specific treatments from the average behavior over time.

To estimate the posterior distribution of the coefficients in equation (B.1), we use JAGS 3.4. We use 4 independent chains. We discard 5000 samples for adaptation and burnin and we use 10000 samples for each of the 4 chains to estimate the coefficients.

Appendix B.1. Contribution behavior

Appendix B.1.1. Non-punishers

Here, we estimate the contribution behavior of non-punishers in each treatment. As we have several observations per group and per subjects (as we have three non-punishers in each group), we estimate equation (B.1) with an additional random effect $\epsilon_{i,k}$, representing the random effect of the individual i within the group k, with $k \in \{1, \ldots, 24\}$.

Similarly, the vague priors are:

$$\begin{gathered} \beta_{0,1,2,3,4} \sim \mathcal{N}(0, .0001) \\ \tau_l \sim \Gamma(\frac{m_l^2}{s_l^2}, \frac{m_l^2}{s_l^2}) \text{ with } m_l \sim \Gamma(1,1)\,,\ s_l \sim \Gamma(1,1) \text{ with } l \in \{1,2,3\} \end{gathered}$$

Table (B.11) shows the mean estimation results of the estimated β of Equation (B.1) estimated for non-punishers by Bayesian methods, with the dependent variable "Contribution". It also shows the 95% credible interval, the probability that $\beta < 0$, the effective sample size (sseff), and the potential scale reduction factor (psrf). Figure (B.5) visualizes the estimated βs with the corresponding 95% credible intervals.

[23] Hence, we assume that the contribution follows a fundamental function over time for all treatments and we want to estimate the deviations from this fundamental function.

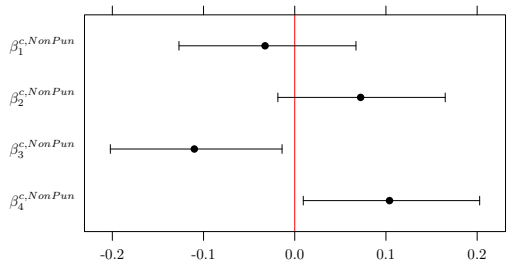

Figure B.5: The graphs show the 95%-credible intervals for the coefficients of the estimation results for Equation (B.1) with the dependent variable: "Contribution to the public-good of non-punishers".

	Mean	CI_{95}	$P(\beta) < 0)$	SSeff	Psrf
β_0	-0.088	[-1.34,1.11]	0.469	95	1.034
β_1	-0.032	[-0.13,0.07]	0.267	236	1.024
β_2	0.073	[-0.02,0.16]	0.941	268	1.031
β_3	-0.11	[-0.2,-0.01]	0.013	251	1.006
β_4	0.105	[0.01,0.21]	0.984	243	1.023
$\beta_4 - \beta_3$	0.215	[0.1,0.33]	0	354	1.005
$\beta_4 - \beta_2$	0.032	[-0.08,0.14]	0.278	373	1.003
$\beta_4 - \beta_1$	0.136	[0.02,0.26]	0.012	306	1.003
$\beta_3 - \beta_2$	-0.183	[-0.29,-0.07]	1	359	1.007
$\beta_3 - \beta_1$	-0.079	[-0.19,0.03]	0.918	363	1.005
$\beta_2 - \beta_1$	0.104	[-0.01,0.22]	0.036	354	1.003

Table B.11: Bayesian estimation results for Equation (B.1) with the dependent variable: "Contribution to the public-good of non-punishers".

Similar to section 4.3 it can be seen that transparency leads to a higher contribution of non-punishers over time compared to no-transparency under low power. This effect is smaller under high power.

Appendix B.1.2. Punishers

Here we estimate model B.1 with the dependent variable "contribution to the public-good" estimated for punishers only.

Table (B.12) shows the mean estimation results, the 95% credible interval, the probability that $\beta < 0$, the effective sample size (sseff), and the potential scale reduction factor (psrf). Figure (B.6) visualizes the estimated βs with the corresponding 95% credible intervals.

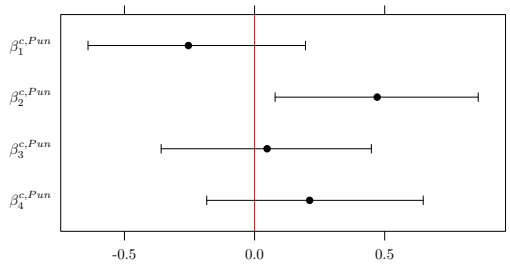

Figure B.6: The graphs show the 95%-credible intervals for the coefficients of the estimation results for Equation (B.1) with the dependent variable: "Contribution to the public-good of punishers".

	Mean	CI_{95}	$P(\beta) < 0)$	SSeff	Psrf
β_0	-1.078	[-4.63,2.22]	0.258	60	1.038
β_1	-0.245	[-0.65,0.18]	0.121	169	1.038
β_2	0.471	[0.1,0.88]	0.987	167	1.016
β_3	0.045	[-0.38,0.43]	0.59	164	1.015
β_4	0.213	[-0.19,0.64]	0.852	166	1.022
$\beta_4 - \beta_3$	0.167	[-0.24,0.57]	0.202	330	1.009
$\beta_4 - \beta_2$	-0.258	[-0.65,0.13]	0.907	339	1.016
$\beta_4 - \beta_1$	0.458	[0.07,0.84]	0.01	342	1.006
$\beta_3 - \beta_2$	-0.426	[-0.83,-0.03]	0.985	323	1.029
$\beta_3 - \beta_1$	0.29	[-0.11,0.68]	0.084	301	1.009
$\beta_2 - \beta_1$	0.716	[0.32,1.11]	0	323	1.036

Table B.12: Bayesian estimation results for Equation (B.1) with the dependent variable: "Contribution to the public-good of punishers".

Similar to section 4.3, the contribution of punishers over time did not significantly increase, and the only treatment under which the contributions increase was the high power, high-transparency treatment.

Appendix B.2. Imposed Norm

As can be seen in Figure 2, the imposed norm does not follow a linear function, and hence we again estimate Equation (B.1) with the dependent variable "imposed norm".

Table (B.13) shows the mean estimation results of the estimated βs, the 95% credible interval, the probability that $\beta < 0$, the effective sample size (sseff), and the potential scale reduction factor (psrf). Figure (B.7) visualizes the estimated βs with the corresponding 95% credible intervals.

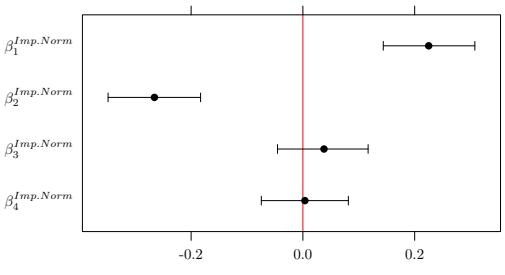

Figure B.7: The graphs show the 95%-credible intervals for the coefficients of the estimation results for Equation (B.1) with the imposed norm as the dependent variable.

	Mean	CI_{95}	$P(\beta) < 0)$	SSeff	Psrf
β_0	0.018	[-1.17,1.25]	0.536	95	1.021
β_1	0.225	[0.14,0.3]	1	376	1.01
β_2	-0.266	[-0.35,-0.18]	0	390	1.01
β_3	0.037	[-0.04,0.12]	0.812	401	1.017
β_4	0.004	[-0.07,0.08]	0.535	409	1.004
$\beta_4 - \beta_3$	-0.033	[-0.14,0.08]	0.724	428	1.014
$\beta_4 - \beta_2$	0.269	[0.16,0.39]	0	418	1.011
$\beta_4 - \beta_1$	-0.221	[-0.33,-0.12]	1	435	1.013
$\beta_3 - \beta_2$	0.302	[0.19,0.41]	0	427	1.014
$\beta_3 - \beta_1$	-0.188	[-0.3,-0.07]	1	435	1.015
$\beta_2 - \beta_1$	-0.491	[-0.61,-0.38]	1	398	1.004

Table B.13: Bayesian estimation results for Equation (B.1) with the imposed norm as the dependent variable.

The results of section 4.2 are replicated using loess splines. It can be seen that, under low transparency, high power leads to higher imposed norms, that transparency does not have much of an effect on the imposed norm under low power, and that transparency leads to lower imposed norms under high power.

Appendix B.3. Abusive behavior

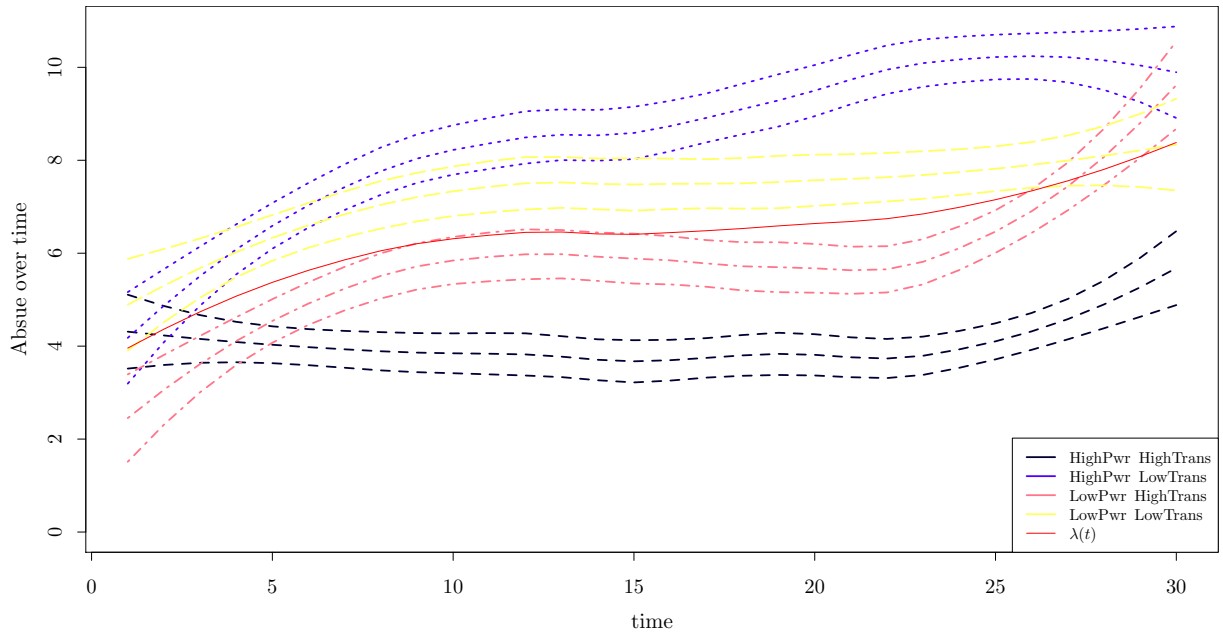

Figure B.8: Splines of the abusive behavior for the treatments with the corresponding confidence interval bands and additionally with λ.

Appendix B.3.1. Abusive behavior defined as in main part of the paper

In this section we replicate the results of section 4.1 with non-linear behavior of abuse over time (as can be seen in Figure B.8).

Table (B.14) shows the mean estimation results of the estimated β of Equation (B.1) with dependent variable "abuse" estimated by Bayesian methods, the 95% credible interval, the probability that $\beta < 0$, the effective sample size (sseff), and the potential scale reduction factor (psrf). Figure (B.9) visualizes the estimated βs with the corresponding 95% credible intervals.

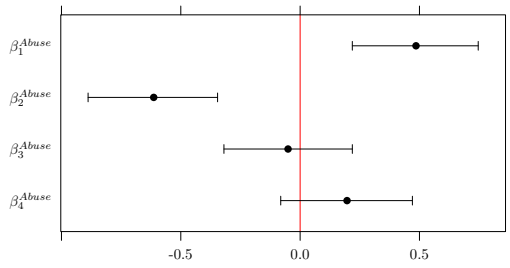

Figure B.9: The graphs show the 95%-credible intervals for the coefficients of the estimation results for Equation (B.1) with the abusive behavior as the dependent variable.

	Mean	CI_{95}	$P(\beta) < 0)$	SSeff	Psrf
β_0	-0.068	[-1.64,1.59]	0.458	206	1.007
β_1	0.485	[0.22,0.75]	1	716	1.005
β_2	-0.612	[-0.88,-0.34]	0	662	1.012
β_3	-0.05	[-0.32,0.22]	0.358	694	1.007
β_4	0.195	[-0.08,0.47]	0.918	691	1.002
$\beta_4 - \beta_3$	0.245	[-0.11,0.6]	0.091	797	1.004
$\beta_4 - \beta_2$	0.807	[0.45,1.18]	0	751	1.009
$\beta_4 - \beta_1$	-0.29	[-0.64,0.04]	0.95	867	1.005
$\beta_3 - \beta_2$	0.562	[0.21,0.9]	0.001	853	1.005
$\beta_3 - \beta_1$	-0.535	[-0.89,-0.19]	0.999	779	1.002
$\beta_2 - \beta_1$	-1.097	[-1.45,-0.74]	1	815	1.008

Table B.14: Bayesian estimation results for Equation (B.1) with the abusive behavior as the dependent variable.

Again all the results of section 4.1 are replicated using loess splines. Transparency has no effect on abusive behavior in the low power setting, power corrupts under low transparency and high transparency reduces abusive behavior in the high power setting.

Appendix B.3.2. Abusive behavior defined as the punishers deviation from the average non-punisher

We also estimate the non-linear behavior of abuse with the alternative definition of abuse.

Table (B.15) shows the mean estimation results of the estimated β of Equation (B.1) with dependent variable "deviation of punishers contribution from the average non-punishers contribution" estimated by Bayesian methods, the 95% credible interval, the probability that $\beta < 0$, the effective sample size (sseff), and the potential scale reduction factor (psrf). Figure (B.10) visualizes the estimated βs with the corresponding 95% credible intervals.

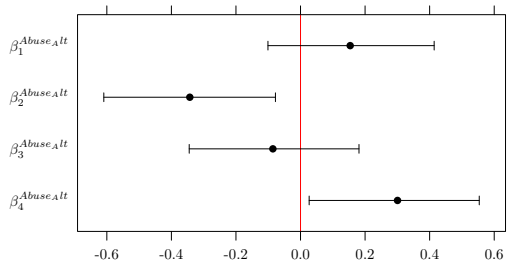

Figure B.10: The graphs show the 95%-credible intervals for the coefficients of the estimation results for Equation (B.1) with the dependent variable: "deviation of punishers contribution from the average non-punishers contribution".

	Mean	CI_{95}	$P(\beta) < 0)$	SSeff	Psrf
β_0	-0.125	[-1.86,1.5]	0.453	246	1.004
β_1	0.155	[-0.1,0.41]	0.88	831	1.005
β_2	-0.342	[-0.61,-0.08]	0.006	771	1.001
β_3	-0.084	[-0.35,0.18]	0.268	789	1.001
β_4	0.298	[0.03,0.55]	0.983	794	1.005
$\beta_4 - \beta_3$	0.382	[0.02,0.73]	0.02	890	1.002
$\beta_4 - \beta_2$	0.64	[0.29,0.99]	0	865	1.005
$\beta_4 - \beta_1$	0.143	[-0.21,0.47]	0.207	922	1.003
$\beta_3 - \beta_2$	0.258	[-0.09,0.62]	0.073	900	1.002
$\beta_3 - \beta_1$	-0.239	[-0.59,0.11]	0.911	886	1.003
$\beta_2 - \beta_1$	-0.497	[-0.83,-0.15]	0.998	930	1.004

Table B.15: Bayesian estimation results for Equation (B.1) with the dependent variable: "deviation of punishers contribution from the average non-punishers contribution".

And here again all the results of section Appendix A.1 are replicated. Hence, all results are robust to the assumption of linearity.