

Hoelt, Leonard; Mill, Wladislaw

Working Paper

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

Preprints of the Max Planck Institute for Research on Collective Goods, No. 2017/15

Provided in Cooperation with:

Max Planck Institute for Research on Collective Goods

Suggested Citation: Hoelt, Leonard; Mill, Wladislaw (2017) : Abuse of power: An experimental investigation of the effects of power and transparency on centralized punishment, Preprints of the Max Planck Institute for Research on Collective Goods, No. 2017/15, Max Planck Institute for Research on Collective Goods, Bonn

This Version is available at:

<https://hdl.handle.net/10419/174494>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



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

Leonard Hoefl
Wladislaw Mill





Abuse of Power

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

Leonard Hoefl / Wladislaw Mill

August 2017

Abuse of Power

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

Leonard Hoef^{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 Hoef), wladislaw.mill@uni-jena.de (Wladislaw Mill)

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

35 2. Literature

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

40 Peer punishment may, however, fall short of the desired goal. It can be highly inefficient when punishers
41 fail to coordinate and therefore punish too little or too much. This is especially pertinent as each would prefer
42 the other to bear the costs of punishing, which could lead to a failure to coordinate so that no punishment
43 is meted out at all. Furthermore, the external validity of peer punishment remains uncertain (Guala, 2012).
44 All but small scale, close-knit societies rely on some form of institutional punishment. Empirical studies
45 could show that participants exhibit a preference for such institutions (Zhang et al., 2014, Traulsen et al.,
46 2012, Hilbe et al., 2013), even being willing to pay for them (Andreoni and Gee, 2012). If institutions
47 are available, participants can organize to implement them efficiently (Putterman et al., 2011) and their
48 effectiveness depends on whether they are endogenously chosen (Markussen et al., 2014). This strain of
49 literature models institutions as a centralized and automatic punishment mechanism: It is exogenously
50 determined and fixed. In real societies, the institutions depend on members of society themselves and are
51 therefore not only endogenously implemented, but depend on the strategic choices of their members.

52 To see how participants use institutional power, experiments investigated how subjects would fill roles of
53 designated punishment. If only one person is endowed with punishment power in a public-goods-game as a
54 second-party (O’Gorman et al., 2009) or third-party punisher (Baldassarri and Grossman, 2011b), the social
55 dilemma is mitigated, even at a personal cost. This held even when strategic punishment was excluded, as
56 the second-party punisher was randomly rematched every round and the third party punisher did not benefit
57 from group contributions. Third-party punishers actively promote their own social preferences (Engel and
58 Zhurakhovska, 2017). While no design targets the goals of single second-party punishers, some studies shed
59 light on those of second- and third-party peer punishers with conflicting results: Carpenter and Matthews
60 (2009) find that second-party punishers enforce absolute and fixed norms, while third-party punishers enforce
61 norms relative to group contributions, yet in Carpenter and Matthews (2012) second-party punishers enforce
62 conformism and third parties fixed minimum contribution norms. According to Zhou et al. (2017), third
63 parties punish more frequently, severely and less antisocially. Motivations considered for deciding to punish
64 reach from fairness norms (Falk et al., 2005), egalitarian motives (Johnson et al., 2009, Leibbrandt and
65 López-Pérez, 2012), reputation gain and leadership (O’Gorman et al., 2009) to destructive impulses such as
66 spite and retaliation (Herrmann et al., 2008, Houser and Xiao, 2010, Falk et al., 2005). Participants seem
67 to understand that there are those who punish prosocially: Fehr and Williams (2013) show that subjects
68 manage to elect cooperative leaders into positions of power, and cooperative non-punishers are willing to
69 empower those who punish in the interest of the group (Gross et al., 2016). Although there are circumstances
70 where decentralized peer punishment is preferred, under imperfect information third-party punishers (who,
71 however, benefit from contributions) are favored (Nicklisch et al., 2016).

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

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

89 In addition to the literature on punishment and its motivations, our study also contributes to informa-
90 tional feedback in relation to punishment. In our design, only the information of the non-punishers is varied:
91 In the high information treatment, the non-punishers gain additional feedback on individual contributions,
92 in the low information treatment only on aggregates. Absent of punishment, this leads to no contribution
93 differences (Croson, 2001). Although studies suggest that if punishers themselves are subject to noise in feed-
94 back of others behavior the cooperation is more difficult to sustain (Patel et al., 2010, Ambrus and Greiner,
95 2012, Fischer et al., 2016) this does not necessarily shed light on how non-punishers react to informational
96 differences. As transparency in public-goods with heterogeneous action spaces improves contributions only
97 when punishment exists (Khadjavi et al., 2014), we might at least suspect that more information does not
98 always benefit contributions when some inequality exists in society.

99 In our design, rational choice theory, assuming selfish and money-maximizing participants, would predict
100 that every participant unequivocally defects and punishers do not use their punishment power. There is no
101 immediate way to form predictions based on the previous experimental evidence, as the possibility of abuse
102 we are interested in is only generated by our new design, in which a singular second-party punisher holds all
103 punishment power. Therefore, we test straightforward and publicly held intuitions about the relationship
104 between power and transparency.

105 We hypothesize that:

106 **H1** *Higher power leads to more abuse.*

107 **H2** *Transparency leads to less abuse.*

108 **H3** *Higher power impedes the effect of transparency.*

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

111 3. Materials and Methods

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

117 3.1. Measurements

118 3.1.1. Public-goods-game task

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

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

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

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

128 In the second stage, only the punisher (who was referred to as “ D ”) would be informed about the
 129 first-stage contributions of all group members. The participants were shown in random order each round
 130 anew to rule out reputation effects from previous rounds. D (the punisher) was now asked to indicate how
 131 much he would punish subject i (ζ_i , $i \neq D$)². For this purpose he was equipped with 30 tokens in the
 132 “low” power treatment. In the high-power treatment, the punisher was equipped with 120 tokens. We set
 133 the low-power treatment to have enough punishment points to deter every participant from free-riding³ to
 134 eliminate unobservable strategic considerations and focus on a purely behavioral effect. Each token could
 135 be used by the punisher to deduct a targeted subject’s payoff by one token. Unused tokens were not added
 136 to the payoff of D to rule out equality concerns,⁴ so the contributions of the punisher could be compared to
 137 the contributions of others directly. The other three group members were just shown a blank screen asking
 138 them to wait for the decision of the punisher. The payoff π_i of the participant $i \neq D$ can therefore be
 139 formalized in the following way (the payoff of the punisher is described by equation (1)):

$$\pi_i = 20 - c_i + 0.5 \cdot \sum_{j \in \{1, n\}} c_j - \zeta_i \quad (2)$$

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

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

151 3.1.2. Additional measurements

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

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

²To avoid framing and demand effects, we referred to the act as “reducing the payoff”.

³Note 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.

⁴In 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.

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

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

161 *3.2. Participants and Procedure*

162 384 participants (53 % female) were recruited with the online registration software Hroot (Bock et al.,
163 2014). The experiment was conducted at the BonnEconLab and consisted of 16 sessions each with 24
164 participants. The participants' age ranged from 16 to 57 years (Mdn = 23). Most students were bachelor
165 students (Semester Mdn=5). The average earning was 14.54 € (including a 4 € show-up fee) and the
166 experiment lasted 1.5 hours (including setting, video instructions, payoff etc.). All measurements were
167 computerized with the experimental software z-Tree (Fischbacher, 2007).

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

173 **4. Data Analysis**

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

180 *4.1. Abuse of Punishment Power*

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

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

⁵Hence, 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).

⁶The 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.

⁷In 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.

⁸Note 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.

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

197 Over all treatments, 83.3 % of subjects abused their power at least once. Table 1 reports this percentage
 198 for each of the treatments. The table also reports the average percentage of rounds in which subjects behave
 199 abusively, given that they behave abusively at least once in the whole experiment. High transparency does
 200 curb abuse in the high-power treatment, $t(45.7) = 2.5$, $p = 0.02$; however, it has no significant effect on the
 201 average frequency of abusive behavior under low power $t(45.6) = 0.8$, $p \geq 0.05$. As can be seen in Figure 1,
 202 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

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

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

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

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

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

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

218 **Result 1 b** Abuse increased over time.

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

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

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

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

¹⁰Note 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.

¹¹Note: 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.

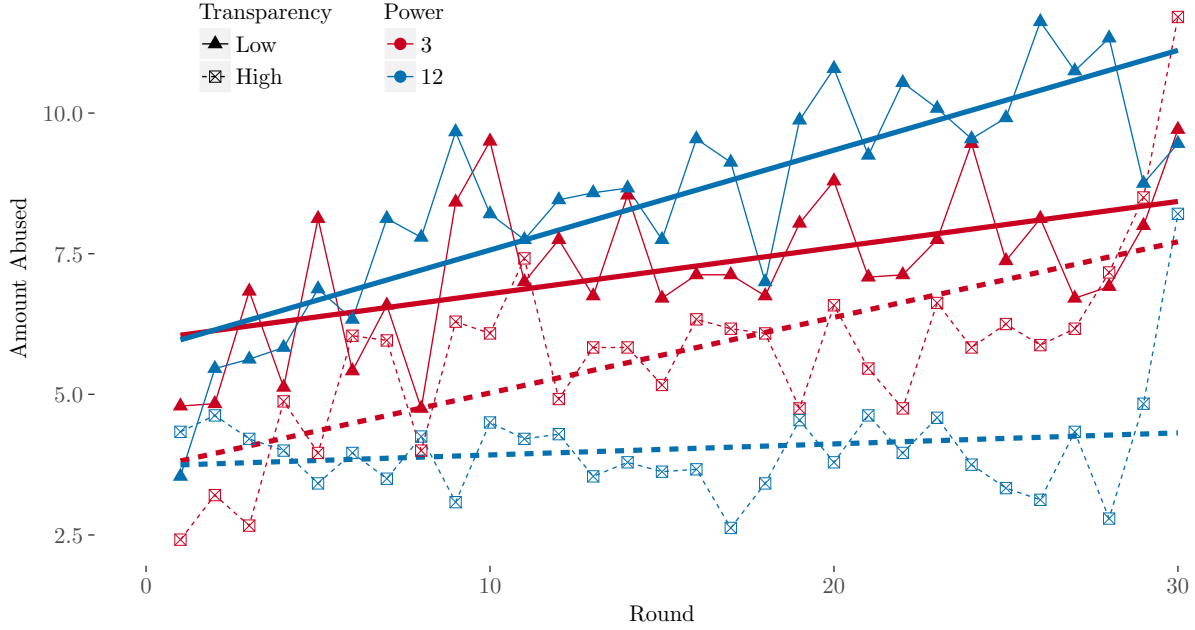


Figure 1: Amount abused in each of the four treatments over time with additional linear regression lines. Blue lines represent the punisher’s abusive behavior in the high-power treatments, while red lines represent the low-power treatments. The high-transparency treatments are represented by dashed lines 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			
	<i>Abuse Amount</i>		<i>High Power</i>	<i>Low Power</i>
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)		
<i>t</i>	0.07*** (0.02)	0.08 (0.05)	0.17*** (0.02)	0.06** (0.02)
<i>t</i> x HighTrans	0.05 (0.03)	0.06* (0.03)	-0.16*** (0.02)	0.05 (0.03)
<i>t</i> x HighPwr	0.10*** (0.03)	0.09** (0.03)		
<i>t</i> 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: 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.

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

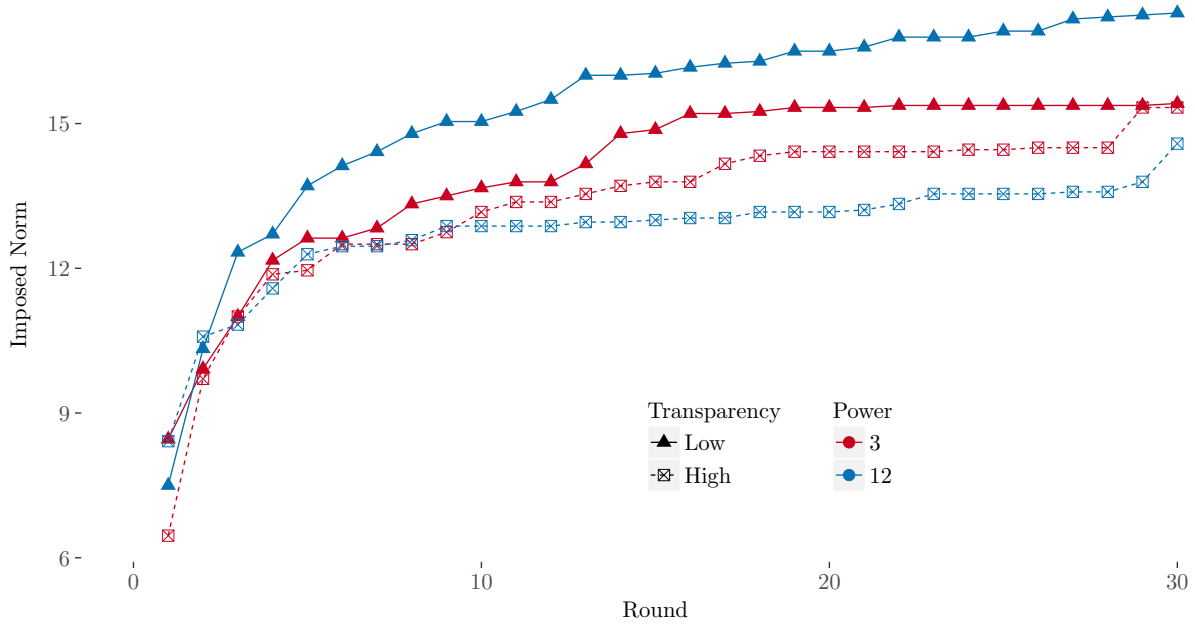


Figure 2: Imposed norms over time in each of the four treatments.

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.

227 *4.2. Imposed Norm*

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

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

234 Using a linear regression¹³ for the imposed norm with fixed effects of treatments, time, the interaction
 235 of both and controlling for subject-specific random effects, we see in Table 3 that high power leads to a
 236 stronger increase in the imposed norm, although this effect is reversed if transparency is high.

237 All results are robust to the inclusion of controls.

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

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

¹²Here, 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

¹³Note: 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:	·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 $M=9.38$ ($SD^{14}=3.57$) while non-punishers contributed $M=16.19$ ($SD=4.28$) 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 than non-punishers.

¹⁴Average of the standard deviations over all rounds.

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

¹⁶Note 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.

¹⁷We 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.

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

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

267 **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)
$t \times$ HighTrans	0.09*** (0.01)	0.09*** (0.01)	0.01 (0.03)	0.01 (0.03)
$t \times$ HighPwr	0.01 (0.01)	0.01 (0.01)	-0.03 (0.03)	-0.02 (0.03)
$t \times$ 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.

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

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

273 5. Discussion

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

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

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

299 Acknowledgements

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

305 References

- 306 Ambrus, A., Greiner, B., 2012. Imperfect public monitoring with costly punishment: an experimental study. *American Economic*
307 *Review* 102 (7), 3317–32.
- 308 Andreoni, J., Croson, R., 2008. Partners versus strangers: Random rematching in public goods experiments. In: Plott, C. R.,
309 Smith, V. L. (Eds.), *Handbook of Experimental Economics Results*, 1st Edition. Vol. 1, Part 6. Elsevier, Ch. 82, pp. 776–783.
- 310 Andreoni, J., Gee, L., 2012. Gun for hire: Delegated enforcement and peer punishment in public goods provision. *Journal of*
311 *Public Economics* 96 (11), 1036–1046.
- 312 Azrieli, Y., Chambers, C. P., Healy, P. J., 2015. Incentives in experiments: A theoretical analysis, mimeo.
- 313 Back, M. D., Kuefner, A. C. P., Dufner, M., Gerlach, T. M., Rauthmann, J. F., Denissen, J. J. A., 2013. Narcissistic admiration
314 and rivalry: Disentangling the bright and dark sides of narcissism. *Journal of Personality and Social Psychology* 105 (10),
315 1013–1037.
- 316 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.
- 317 Baldassarri, D., Grossman, G., 2011b. Centralized sanctioning and legitimate authority promote cooperation in humans.
318 *Proceedings of the National Academy of Sciences* 108, 11023–11027.
- 319 Bock, O., Baetge, I., Nicklisch, A., 2014. hroot: Hamburg registration and organization online tool. *European Economic Review*
320 71 (C), 117–120.
- 321 Carpenter, J., Matthews, P. H., 2009. What norms trigger punishment? *Experimental Economics* 12 (3), 272–288.
- 322 Carpenter, J. P., Matthews, P. H., 2012. Norm enforcement: anger, indignation, or reciprocity? *Journal of the European*
323 *Economic Association* 10 (3), 555–572.
- 324 Cinyabuguma, M., Page, T., Putterman, L., 2006. Can second-order punishment deter perverse punishment? *Experimental*
325 *Economics* 9 (3), 265–279.
- 326 Croson, R. T., 2001. Feedback in voluntary contribution mechanisms: An experiment in team production. In: Isaac, R. M.,
327 Norton, D. A. (Eds.), *Research in Experimental Economics*. Vol. 8 of *Research in Experimental Economics*. Emerald Group
328 Publishing Limited, pp. 85–97.
- 329 Engel, C., Zhurakhovska, L., 2017. You are in charge: Experimentally testing the motivating power of holding a judicial office.
330 *The Journal of Legal Studies* 46 (1), 1–50.
- 331 Falk, A., Fehr, E., Fischbacher, U., 2005. Driving forces behind informal sanctions. *Econometrica* 73 (6), 2017–2030.
- 332 Fehr, E., Fischbacher, U., 2004. Third-party punishment and social norms. *Evolution and Human Behavior* 25 (2), 63–87.
- 333 Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. *Nature* 415 (6868), 137–140.
- 334 Fehr, E., Williams, T., 2013. Endogenous emergence of institutions to sustain cooperation. Tech. rep., Working Paper.
- 335 Fischbacher, U., 2007. z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10 (2), 171–178.
- 336 Fischer, S., Grechenig, K., Meier, N., 2016. Monopolizing sanctioning power under noise eliminates perverse punishment but
337 does not increase cooperation. *Frontiers in Behavioral Neuroscience* 10.
- 338 Gintis, H., 2016. *Individuality and entanglement: the moral and material bases of social life*. Princeton University Press.
- 339 Gross, J., Mder, Z. Z., Okamoto-Barth, S., Riedl, A., 2016. Building the leviathan - voluntary centralisation of punishment
340 power sustains cooperation in humans. *Scientific Reports* 6, 20767.
- 341 Guala, F., 2012. Reciprocity: Weak or strong? what punishment experiments do demonstrate. *Behavioral and Brain Sciences*
342 35 (1), 1–15.
- 343 Herrmann, B., Thöni, C., Gächter, S., 2008. Antisocial punishment across societies. *Science* 319 (5868), 1362–1367.
- 344 Hilbe, C., Traulsen, A., Röhl, T., Milinski, M., 2013. Democratic decisions establish stable authorities that overcome the
345 paradox of second-order punishment. *Proceedings of the National Academy of Sciences* 111 (2), 201315273–756.
- 346 Hoefft, L., Mill, W., 2017. Selfish punishers. *Economic Letters* 157, 41–44.
- 347 Houser, D., Xiao, E., 2010. Inequality-seeking punishment. *Economics Letters* 109 (1), 20–23.
- 348 Johnson, T., Dawes, C. T., Fowler, J. H., McElreath, R., Smirnov, O., 2009. The role of egalitarian motives in altruistic
349 punishment. *Economics Letters* 102 (3), 192–194.
- 350

- 351 Khadjavi, M., Lange, A., Nicklisch, A., 2014. The social value of transparency and accountability: experimental evidence from
352 asymmetric public goods games. *WiSo-HH Working Paper 1* (12).
- 353 Leibbrandt, A., López-Pérez, R., 2012. An exploration of third and second party punishment in ten simple games. *Journal of*
354 *Economic Behavior & Organization* 84 (3), 753–766.
- 355 Marcus, D. K., Zeigler-Hill, V., Mercer, S. H., Norris, A. L., 2014. The psychology of spite and the measurement of spitefulness.
356 *Psychological Assessment* 26 (2), 563–574.
- 357 Markussen, T., Putterman, L., Tyran, J.-R., 2014. Self-organization for collective action: An experimental study of voting on
358 sanction regimes. *The Review of Economic Studies* 81 (1), 301–324.
- 359 Murphy, R. O., Ackerman, K. A., 2014. Social value orientation: Theoretical and measurement issues in the study of social
360 preferences. *Personality and Social Psychology Review* 18 (1), 13–41.
- 361 Murphy, R. O., Ackerman, K. A., Handgraaf, M. J. J., 2011. Measuring social value orientation. *Judgment and Decision Making*
362 6 (8), 771–781.
- 363 Nicklisch, A., Grechenig, K., Thöni, C., 2016. Information-sensitive Leviathans. *Journal of Public Economics* 144, 1–13.
- 364 Nikiforakis, N., 2010. Feedback, punishment and cooperation in public good experiments. *Games and Economic Behavior*
365 68 (2), 689–702.
- 366 Nikiforakis, N., Normann, H.-T., 2008. A comparative statics analysis of punishment in public-good experiments. *Experimental*
367 *Economics* 11 (4), 358–369.
- 368 O’Gorman, R., Henrich, J., Van Vugt, M., 2009. Constraining free riding in public goods games: designated solitary punishers
369 can sustain human cooperation. *Proceedings of the Royal Society of London B: Biological Sciences* 276 (1655), 323–329.
- 370 Patel, A., Cartwright, E., van Vugt, M., 2010. Punishment cannot sustain cooperation in a public good game with free-rider
371 anonymity, mimeo.
- 372 Putterman, L., Tyran, J.-R., Kamei, K., 2011. Public goods and voting on formal sanction schemes. *Journal of Public Economics*
373 95 (910), 1213–1222.
- 374 Ramalingam, A., Rojo Arjona, D., Schram, A., Van Leeuwen, B., 2015. Authority and Centrality: Power and Cooperation in
375 Social Dilemma Networks. *IAST Working Papers 15-23*, Institute for Advanced Study in Toulouse (IAST).
- 376 Reuben, E., Riedl, A., 2013. Enforcement of contribution norms in public good games with heterogeneous populations. *Games*
377 *and Economic Behavior* 77 (1), 122–137.
- 378 Traulsen, A., Röhl, T., Milinski, M., 2012. An economic experiment reveals that humans prefer pool punishment to maintain
379 the commons. *Proceedings of the Royal Society of London B: Biological Sciences* 279 (1743), 3716–3721.
- 380 Xiao, E., 2013. Profit-seeking punishment corrupts norm obedience. *Games and Economic Behavior* 77 (1), 321–344.
- 381 Xiao, E., Tan, F., 2014. Justification and legitimate punishment. *Journal of Institutional and Theoretical Economics* 170 (1),
382 168–188.
- 383 Zhang, B., Li, C., Silva, H., Bednarik, P., Sigmund, K., 2014. The evolution of sanctioning institutions: an experimental
384 approach to the social contract. *Experimental Economics* 17 (2), 285–303.
- 385 Zhou, Y., Jiao, P., Zhang, Q., 2017. Second-party and third-party punishment in a public goods experiment. *Applied Economics*
386 *Letters* 24 (1), 54–57.

387 **Appendix A. Additional regressions**

388 *Appendix A.1. Alternative definition of abuse*

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

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

393 It is worth pointing out several shortcomings of this approach.

- 394 • This alternative approach does not take punishment into account. Hence, subjects are considered
395 abusive even if they have never forced any other group member to contribute anything to the public-
396 good (in case they contribute less than the average non-punisher). Similarly, subjects who do punish the
397 other group members are not considered abusive if the non-punishers are not sensitive to punishment
398 and contribute more than or equal to the punisher. In these instances, the average group contribution
399 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 linear¹⁸ 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)
<i>t</i>	0.08*** (0.02)	0.08 (0.05)
<i>t</i> x HighTrans	0.09** (0.03)	0.09** (0.03)
<i>t</i> x HighPwr	0.05 (0.03)	0.04 (0.03)
<i>t</i> 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: 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 punisher¹⁹ 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) ($\chi^2(0) = 11156.418$, $p \leq 0.01$).

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

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

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

427 We use the following econometric model to estimate punishment behavior:

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

428 where HighPwr is a dummy with value one for the high-power treatment and zero otherwise. HighTrans
429 is a dummy with value one if the subject was in the high-transparency treatment and zero otherwise. $\overline{c_{i \neq D,k,t}}$
430 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
431 the maximum of 0 and the deviation of the contribution of subject i from the average contribution of the
432 non-punishers. ϵ_i and $\epsilon_{i,k}$ are the level 1 and level 2 random effects of groups and individuals.

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

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

439 **Result 4 a** The deviation from the average contribution of non-punishers was punished harshly and in-
440 creased over time.

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

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

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

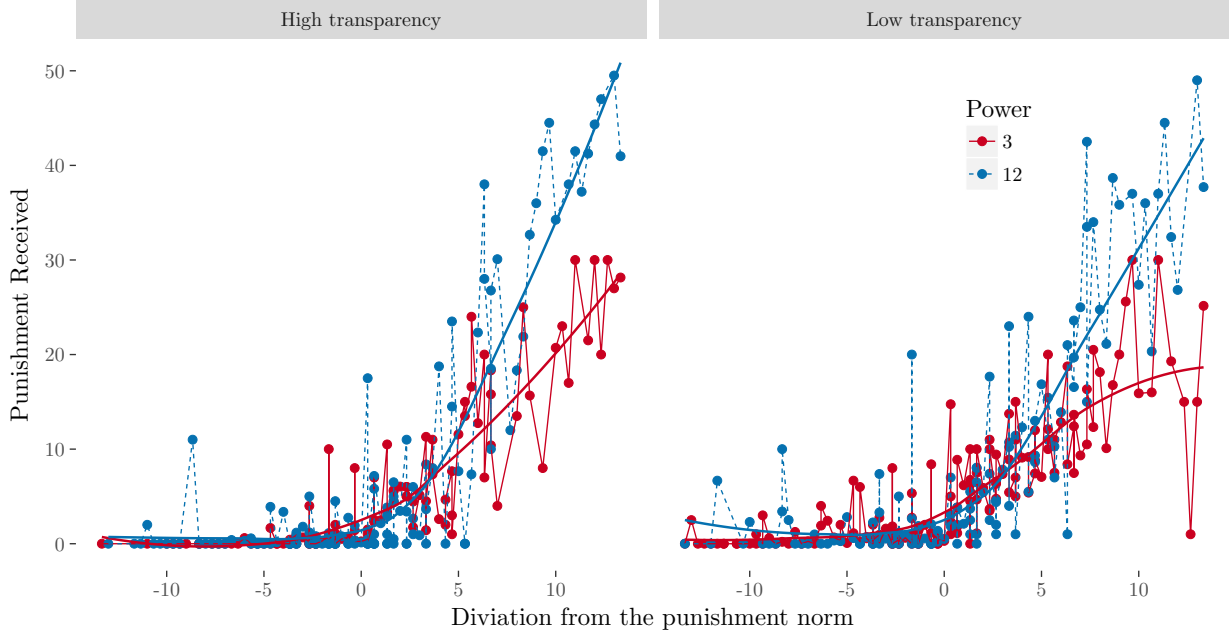


Figure A.4: Punishment received for deviating from the average contribution of the non-punishers with loess splines. Blue lines represent the received punishment in the high-power treatments, while red lines denote the low-power treatments. The graph on the right shows the received punishment in the 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)
$\bar{c}_N - c_j$	1.57*** (0.06)
$\bar{c}_N - c_j$ x HighPwr	1.11*** (0.09)
$\bar{c}_N - c_j$ x HighTrans	0.54*** (0.09)
$\bar{c}_N - c_j$ 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 $\bar{c}_N - c_j$	0.01*** (0.003)
t x $\bar{c}_N - c_j$ x HighPwr	-0.005 (0.01)
t x $\bar{c}_N - c_j$ x HighTrans	-0.02*** (0.01)
t x $\bar{c}_N - c_j$ 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 subjects' contribution deviation from the average contribution of all non-punishers.

445 Appendix A.3. Punishment Norms

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

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

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

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

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

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

$$v_{i,k,t} = 1 \text{ if } v_{i,k,t}^* > 0$$

$$p_{i,k,t} = p_{i,k,t}^* \cdot v_{i,k,t}$$

462 where $(x)^+ = \max(x, 0)$. The norm to enforce is γ^v . The norm on how much to punish is γ^p . HighPwr
463 is a dummy with value one if the treatment is the high-power treatment, and zero otherwise. HighTrans
464 is a dummy with value one if the treatment is the high-transparency treatment, and zero otherwise. t is
465 the period. $v_{i,k,t}$ is a variable indicating whether subject i in group k was punished in period t . $p_{i,k,t}$
466 is a variable indicating how much subject i in group k was punished in period t . $c_{i,k,t}$ denotes how much
467 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
468 random effects of groups and individuals. $g(\cdot): (0, 1) \rightarrow \mathbb{R}$ is the link function. We will use the logit link:
469 $g(\mu) = \log(\mu/(1 - \mu))$, as this is easier to interpret.

²¹Note: 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.

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

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

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

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

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

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

	LowPwr \wedge LowTrans		LowPwr \wedge HighTrans		HighPwr \wedge LowTrans		HighPwr \wedge 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 likelihoods of the equation A.3 and A.2 with the specified punishment norms

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

²²Note 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*** (1.07)	-3.55*** (0.60)	-1.54*** (0.35)	-2.83*** (0.48)	0.03 (0.03)	0.15*** (0.04)	0.07* (0.03)	0.22*** (0.05)
$(\gamma^* - c_{i,k,t})^+$	0.90*** (0.08)	0.45*** (0.04)	0.93*** (0.11)	0.42*** (0.03)	0.03*** (0.003)	0.05*** (0.004)	0.02*** (0.003)	0.03*** (0.004)
t	-0.06** (0.02)	-0.11*** (0.02)	-0.07*** (0.01)	-0.07*** (0.01)	-0.001 (0.001)	0.001 (0.001)	-0.0002 (0.001)	0.002 (0.001)
$t \times (\gamma^* - c_{i,k,t})^+$	0.01** (0.003)	0.02*** (0.003)	0.01 (0.01)	-0.0004 (0.002)	0.001*** (0.0002)	-0.0003 (0.0003)	0.001*** (0.0001)	0.001* (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)
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)
$\bar{c}_N - c_j$	0.40*** (0.04)	0.03*** (0.003)
$\bar{c}_N - c_j \times \text{HighPwr}$	0.54*** (0.12)	-0.01** (0.004)
$\bar{c}_N - c_j \times \text{HighTrans}$	0.04 (0.05)	0.02*** (0.01)
$\bar{c}_N - c_j \times \text{HighPwr} \times \text{HighTrans}$	-0.26 (0.14)	-0.01* (0.01)
t	-0.08*** (0.02)	0.002* (0.001)
$t \times \text{HighPwr}$	-0.01 (0.02)	-0.002 (0.002)
$t \times \text{HighTrans}$	-0.03 (0.02)	-0.002 (0.002)
$t \times \text{HighPwr} \times \text{HighTrans}$	0.03 (0.03)	0.001 (0.002)
$t \times \bar{c}_N - c_j$	0.0001 (0.002)	0.0004* (0.0002)
$t \times \bar{c}_N - c_j \times \text{HighPwr}$	0.02* (0.01)	0.0000 (0.0003)
$t \times \bar{c}_N - c_j \times \text{HighTrans}$	0.02*** (0.003)	-0.001* (0.0003)
$t \times \bar{c}_N - c_j \times \text{HighPwr} \times \text{HighTrans}$	-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.

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

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

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

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

501 *Appendix A.4. Imposed Norm*

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

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

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

	Imposed Norm			
	$t \in \{1, \dots, 10\}$	$t \in \{11, \dots, 30\}$	$t \in \{1, \dots, 10\}$	$t \in \{11, \dots, 30\}$
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)
$t \times$ HighTrans	0.04 (0.08)	0.01 (0.01)	0.09 (0.08)	-0.01 (0.01)
$t \times$ HighPwr	0.20** (0.08)	0.02 (0.01)	0.22** (0.08)	0.02 (0.01)
$t \times$ 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: 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.

512 **Appendix B. Bayesian Approaches**

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

517 loess spline for the respective dependent variables over time (average for all treatments)²³, which we call
 518 $\lambda(t)$.

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

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

521 $Ind_{i,t}$ represents the dependent variable (contribution, abuse, etc) of subjects i in round t with $i \in$
 522 $\{1, \dots, n\}$ and $t \in \{0, \dots, 30\}$.

523 $\mathbb{1}_{\text{HighPwr} \wedge \text{LowTrans}}$ denotes a dummy variable which is one if the subject was in the high-power, low-
 524 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
 525 defined accordingly.

526 To account for the nested structure of the data we included ϵ_i as the random effects of the individual i .
 527 $\epsilon_{i,k,t}$ is the residuals.

As vague priors we used:

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

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

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

533 *Appendix B.1. Contribution behavior*

534 *Appendix B.1.1. Non-punishers*

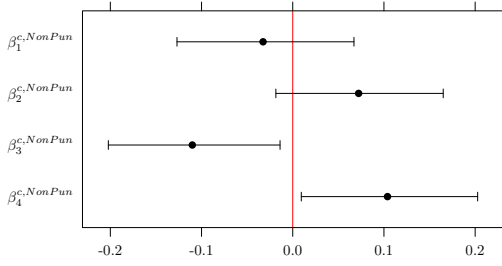
535 Here, we estimate the contribution behavior of non-punishers in each treatment. As we have several
 536 observations per group and per subjects (as we have three non-punishers in each group), we estimate
 537 equation (B.1) with an additional random effect $\epsilon_{i,k}$, representing the random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

539 Similarly, the vague priors are:

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

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

²³Hence, 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.



	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

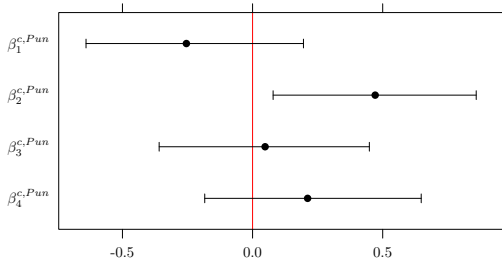
Figure B.5: The graphs show the 95%-credible intervals for the coefficients of the estimation results for Equation (B.1) with the dependent variable: “Contribution to the public-good of non-punishers”. Table B.11: Bayesian estimation results for Equation (B.1) with the dependent variable: “Contribution to the public-good of non-punishers”.

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

546 Appendix B.1.2. Punishers

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

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



	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

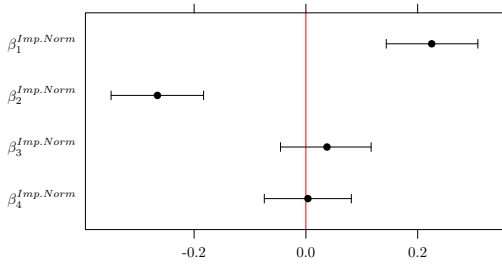
Figure B.6: The graphs show the 95%-credible intervals for the coefficients of the estimation results for Equation (B.1) with the dependent variable: “Contribution to the public-good of punishers”. Table B.12: Bayesian estimation results for Equation (B.1) with the dependent variable: “Contribution to the public-good of punishers”.

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

554 Appendix B.2. Imposed Norm

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

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



	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

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

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

563 *Appendix B.3. Abusive behavior*

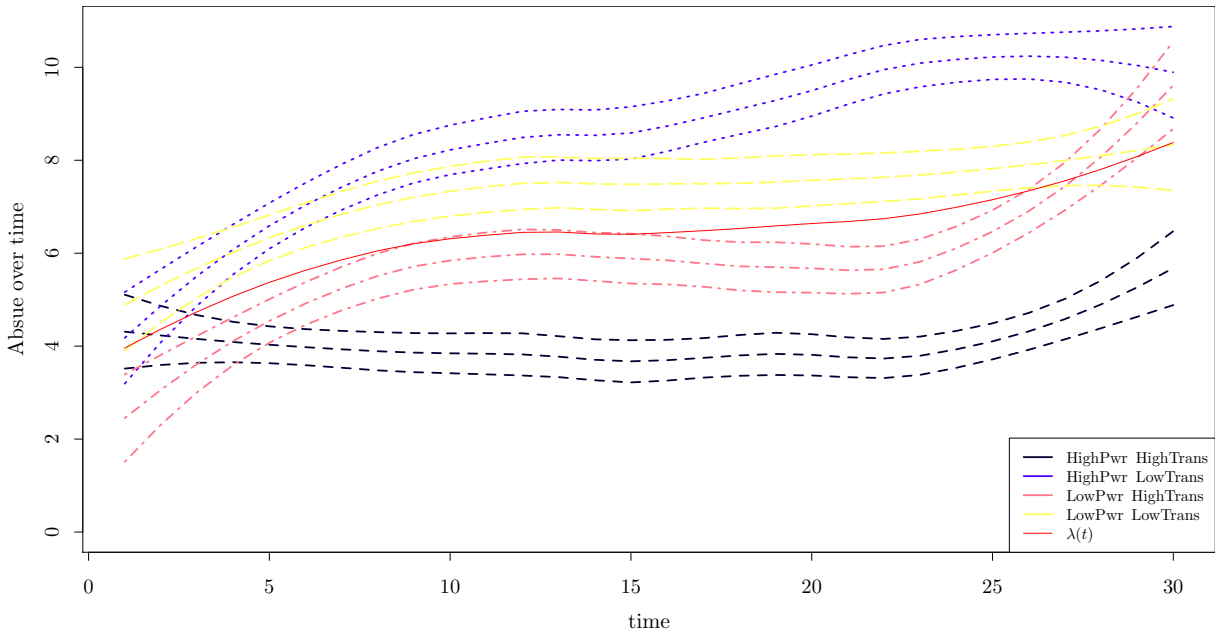
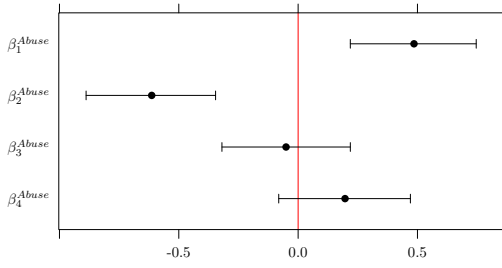


Figure B.8: Splines of the abusive behavior for the treatments with the corresponding confidence interval bands and additionally with λ .

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

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

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



	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

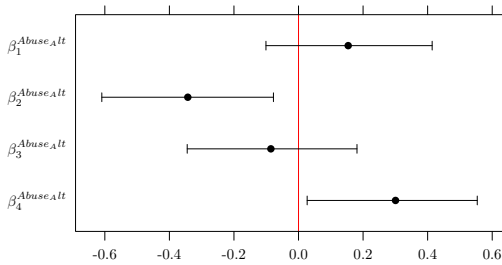
Figure B.9: The graphs show the 95%-credible intervals for the coefficients of the estimation results for Equation (B.1) with the abusive behavior as the dependent variable. Table B.14: Bayesian estimation results for Equation (B.1) with the abusive behavior as the dependent variable.

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

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

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

576 Table (B.15) shows the mean estimation results of the estimated β of Equation (B.1) with dependent
 577 variable “deviation of punishers contribution from the average non-punishers contribution” estimated by
 578 Bayesian methods, the 95% credible interval, the probability that $\beta < 0$, the effective sample size (sseff), and
 579 the potential scale reduction factor (psrf). Figure (B.10) visualizes the estimated β s with the corresponding
 580 95% credible intervals.



	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

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

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