Journal of Economic Behavior and Organization 220 (2024) 305–324

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Research Paper

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

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

[a] Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Straße 10, 53113 Bonn, Germany
[b] University of Mannheim, Department of Economics, L7 3-5, 68131 Mannheim, Germany

ARTICLE INFO

JEL classification:
H41
C92
K42

Keywords:
Power
Corruption
Hypocrisy
Punishment
Transparency
Public-goods game
Designated punishment

ABSTRACT

Punishment institutions are a major guarantor of prosocial behavior. At the same time, their asymmetrical power structure may lead to antisocial behavior itself. We investigate power abuse, understood as the use of power for personal gain, of a single punisher in a public-goods game subject to variations in punishment power and contribution transparency. Using a laboratory experiment 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.

1. Introduction

> Evil is nourished and grows by concealment.
>
> Virgil

In April 2016, the Panama Papers revealed how public officials (and other powerful individuals) were able to employ strategies of tax evasion, systematically avoiding their share of contribution to the public goods. This was widely considered abuse of power: Exploiting institutional power for illegitimate private gain. Public officials enforced obligations they did not adhere to themselves. The revelation sparked an old debate: Is there a corrupting effect of power in the absence of checks and balances? How do we guarantee that those who enforce rules actually comply themselves?

☆ We thank Urs Fischbacher, Alexander Vostroknutov, Eugenio Verrina, Erik Kimbrough, Nikos Nikiforakis, Martin Kocher, Ulrike Malmendier, Julia Sasse, Daniel Houser, Werner Güth, Oliver Kirchkamp, Christoph Engel, Marina Schröder, Henrik Orzen, Bernd Irlenbusch, Dirk Sliwka, Wieland Müller, Jean-Robert Tyran, Andreas Diekmann, Wojtek Przepiorka, Roberto Weber, and Jörg Oechssler for helpful comments. We appreciate comments from the participants of the IMEBESS 2019, VFS 2019, Economic Science Association World Conference 2016, the Jena Econ-Seminar, the Bonn Econ-Seminar, the CODEBE Seminar and the Mannheim Econ-Seminar. We gratefully acknowledge funding from the Max Planck Society and the IMPRS-Uncertainty.

* Corresponding author.
E-mail addresses: hoeft@coll.mpg.de (L. Hoeft), mill@uni-mannheim.de (W. Mill).

https://doi.org/10.1016/j.jebo.2024.02.003
Received 6 October 2022; Received in revised form 22 November 2023; Accepted 6 February 2024

Available online 22 February 2024

In this paper, we conduct a controlled laboratory experiment to study these questions. We do so by employing a standard public goods game with four players. To study a situation of power abuse, we elevated one of those players randomly to have punishing powers (i.e., the punisher), while keeping all other strategic and monetary incentives constant between those four players. One distinctive feature of our setting is that the individuals in power remain in their role for the duration of the experiment (see also our short paper Hoeft and Mill, 2017). Thereby, we isolate the effect of power by excluding self-selection, and selection based on merit (which itself might be based on omitted variables). Further, it gives us the opportunity to study behavior in a setting without proper accountability mechanisms, such as peer punishment. In this setting, we define abuse of power as punishing those who contribute more than the punisher themselves. Abuse of power, thus, is the powerful exempting themselves from the rules they enforce upon others (Hoeft et al., 2019).

Our design captures the power discrepancies created by norm-enforcement institutions which modern, large-scale societies rely on. There are plenty of examples of the powerful exempting themselves from the rules they enforce upon others: Police officers often go unpunished for the (illegal) use of force (Wong, 1998), supervisors or managers can force their coworkers to invest in shared projects while skimping themselves (Xu et al., 2015; Vredenburgh and Brender, 1998), doctors and politicians may use their influence and connections to achieve special treatment (Klitzman, 2007; Olken and Pande, 2012). The same is true, albeit on a much larger scale, for authoritarian regimes or dictatorships, where powerful individuals or groups control all institutional power. While checks and balances are the main institutional answer to power abuse, they work imperfectly even in democratic countries subject to the rule of law, or are altogether absent in autocratic regimes. Increasingly, transparency is proposed as a remedy and key to good governance (Hood and Heald, 2006).

We incorporate transparency in our design by varying the visibility of contributions. More specifically, we have a baseline low-transparency treatment where only the overall contribution to the public good is displayed. Thus, individual contributions are not observable by the players in the non-powerful position (i.e., non-punishers). However, in the high-transparency treatment, all individual contributions to the public good are revealed. Therefore, the contribution behavior of the punisher is plainly visible. This institutional change potentially induces a host of behavioral mechanisms, most of which arguably reduce power abuse.

Finally, to study whether transparency is a suitable remedy under varying conditions of power, we exogenously induce the powerful player with “unlimited” power. More specifically, in the baseline low-power treatment, the punisher is equipped with sufficient power to make all players indifferent between full contribution and free-riding. In the high-power treatment, the punisher is equipped with sufficient power to reduce the payoff of all other players to zero in any situation. Thus, the high-power treatment has a substantially higher threat level.

Our experiment reveals a striking number of punishers abusing their power. Over all treatments, more than 80% of punishers abuse their power at least once. Thus, a vast majority of punishers at least once punish a contribution higher than their own – which is typically considered antisocial (Zhou et al., 2017; Albrecht et al., 2018; Faillo et al., 2013) or hypocritical punishment (Carpenter, 2007; Burton-Chellew and Guérin, 2021; Mieth et al., 2021). In line with our predictions, we find that inducing the punisher with more power leads to more abuse of this power. Most interestingly, we find a surprising interaction between power and transparency. While transparency does not help with the abuse of power in the low-power treatment, it substantially reduces abuse of power in the high-power treatment.

These results highlight three main insights. First, elevating some random players into a position of power for an extended period of time will almost ubiquitously lead to an abuse of power. Second, bestowing people in power with exceptional powers will exacerbate this situation. Finally, the effect of transparency on abuse of power is multilayered. While transparency has a strong and reducing effect on power abuse if the powerful are in an extreme position of power, transparency has only a limited effect if power is constrained.

Our paper makes several contributions. Our paper extends the literature on the corrupting effects of power. We show how the combination of selfish behavior and hypocritical moral evaluations leads to the powerful playing by a different set of rules, which in turn compounds inequality. This may explain why the wealthy contribute less to public goods relative to the poor, especially in corrupt environments (Markussen et al., 2021). The normalization of such different rules even leads to the powerful being exempted by others: Those with visible signs of wealth and power are less likely to be reprimanded for rule violations (Fried et al., 2010).

Our paper also contributes to the literature on transparency, the efficacy of which has been subject to debate with mixed empirical evidence (Kosack and Fung, 2014). Our results highlight how transparency might be an effective tool even in situations where accountability (i.e., institutional or counter-punishment of the punisher) is limited. It also showcases how transparency might interact with the extent of power discrepancies, which informs the emerging literature on how transparency can backfire (Khadjavi et al., 2017).

Finally, our paper contributes to the literature on punishment. In most studies, the punishers were limited in the way they could abuse their position. As third parties, they were unable to profit from enforcing behavior in other groups (Fehr and Fischbacher, 2004; Baldassarri and Grossman, 2011). As single punishers, in their own group, they were randomly chosen each round, so that no punisher could exempt himself from enforcement in future rounds (O’Gorman et al., 2009). Thus, even those studies limited the potential of abuse. Either the single punisher was a third party that could not skimp on their own enforced contribution norms, or the punisher was randomly re-chosen each round, so that they could not enforce high contributions despite undercontributing themselves. We contribute to the literature by using a setting where one random peer is elevated to the position of power for the duration of the experiment. Thus, our design allows us to study how, absent checks and balances, power dynamics evolve and fester. Our design also complements research focusing on democratically appointed sanctioning authorities (Castillo and Hamman, 2020; Castillo et al., 2020).

The remainder of the paper is structured as follows: In Section 2, we discuss the relevant literature. Section 3 will explain the design and the measurements of the experiment. In Section 4, we present the results. Section 5 concludes.

2. Literature

Our paper contributes to several strands of literature. In particular, we contribute to the literature on (peer and centralized) punishment, to the literature investigating power dynamics, and the literature studying the effect of transparency.

Punishment literature Holding people accountable is an effective tool to foster cooperative behavior: Giving participants the power to punish each other has been one of the popular and effective remedies for social dilemmas (Ostrom et al., 1992; Fehr and Fischbacher, 2004; Weber et al., 2018; Rustagi et al., 2010). As this kind of punishment provides a second-order public good of enforcement, its motivations are typically benign: They include fairness norms (Falk et al., 2005), egalitarian motives (Johnson et al., 2009; Leibbrandt and López-Pérez, 2012), reputation gain, leadership (O'Gorman et al., 2009) and only rarely destructive impulses such as spite and retaliation (Herrmann et al., 2008; Houser and Xiao, 2010; Mill and Morgan, 2022, 2021; Falk et al., 2005).

As decentralized punishment can miss its target (Herrmann et al., 2008; Cinyabuguma et al., 2006) and lead to destructive revenge cycles (Nikiforakis and Normann, 2008; Nikiforakis et al., 2012), modern societies rely mostly on institutional accountability. Most of the experimental literature models such institutions as automatic punishment mechanisms (Zhang et al., 2014; Traulsen et al., 2012; Hilbe et al., 2013; Andreoni and Gee, 2012; Putterman et al., 2011; Kosfeld and Rustagi, 2015; Markussen et al., 2014; Sutter et al., 2010). This circumvents the central problem of institutionalized power: It has to be entrusted to individual members and is therefore not automatic, but depends on their strategic choices. We contribute to the literature on institutional punishment by having an individual participant personify the punishment institution.

Few other studies have focused on having single punishers fill the role of a punishment institution (Bolle and Vogel, 2011). They show that 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, 2011), the social dilemma is mitigated, even at a personal cost. This holds even though strategic punishment is excluded, as the second-party punisher is randomly rematched every round, and the third-party punisher does not benefit from group contributions. Contrary to our design, these studies limited the potential of abuse. Either the single punisher was a third party that could not skimp on their own enforced contribution norms, or the punisher was randomly re-chosen each round, so that they could not enforce high contributions despite undercontributing themselves.

As reported in our previous paper (Hoeft and Mill, 2017), our baseline confirms the validity of our experimental design as punishers indeed undercontribute while punishing.[1] In this paper, different from Hoeft and Mill (2017), we provide a detailed analysis of how punishers use their power, how changes in power (manipulated through punishment power) affect the abuse of power, and how and when transparency (manipulated through contribution transparency) can curb the abuse of power.

Literature studying the effect of power Institutionalization may avoid the drawbacks of decentralized punishment but creates power discrepancies that can be abused. By having a single participant in charge of punishment over time, we extend the literature studying the effect of conferring power to individuals. While demonstrating the corrupting effect of power has a long tradition in psychology, economic research has mostly investigated the prosocial nature of punishment power. Psychological research highlights that power corrupts, leading to selfishness (Kipnis, 1972; Galinsky et al., 2015; Lammers et al., 2015), cheating (Lammers et al., 2010; Dubois et al., 2015), social distancing (Lammers et al., 2012), distrust (Schilke et al., 2015), disinhibition (Lammers et al., 2015) and therefore violation of social norms (van Kleef et al., 2015), as well as unjust allocations in economic games (Giurge et al., 2021; van Dijk et al., 2004; Bendahan et al., 2015). Importantly, random assignment to power results in moral hypocrisy (Lammers et al., 2010) as powerful individuals often violate or strategically modify their own moral values (Rustichini and Villeval, 2014; Mallucci et al., 2019; Schier et al., 2016), yet harshly judge and punish others' ethical transgressions (Lammers et al., 2010; Wiltermuth and Flynn, 2013; Mooijman et al., 2015). While some studies have investigated punishment for personal gain (Xiao, 2013; Buffat and Senn, 2018), outright abuse of power in the absence of norm violations is easy to monitor and, therefore, rare. In our design, punishment is still directed at its intended target, but results in double standards. So far, only Castillo and Hamman (2020); Castillo et al. (2020) implemented a similar design where democratically appointed sanctioning authorities undercontribute. In contrast to their study, however, we vary transparency and make punishment costless for the punishers to control for strategic considerations.

Literature studying the effect of transparency Transparency is often seen as the key to combating the negative effects of institutional power discrepancies (Hood and Heald, 2006). Its effectiveness, however, remains in doubt. While some find that transparency can help (Molina et al., 2016; Chen and Ganapati, 2023; Cordis and Warren, 2014) – especially for the underprivileged and even in hierarchical and unequal societies (Peisakhin and Pinto, 2010; Peisakhin, 2012) – the overall evidence is mixed (Kosack and Fung, 2014). Its effectiveness seems to depend on a host of factors, such as the type of corrupt behavior addressed (Parra et al., 2021) and the mechanisms enforcing accountability (Bauhr and Grimes, 2014).

Arguably, transparency's primary mechanism is increasing accountability, which deters selfish behavior (Rus et al., 2012) by making hypocritical use of power visible (Kolstad and Wiig, 2009). Previous experiments, for example, demonstrate less corruption

[1] Note that the current paper partially uses the same data as reported in Hoeft and Mill (2017).

if it is harder to hide and there is a risk of being punished (Azfar and Nelson, 2007; Abbink et al., 2002). In these cases, transparency is a necessary condition for the punishment of corruption, and therefore, transparency works through **accountability**.

Accountability may not always be realistic, especially when law enforcement institutions are ineffective (Dasgupta and Radoniqi, 2021) or power discrepancies are as high as in autocratic states. Transparency can curb abuse of power due to psychological mechanisms like creating trust (Kochel and Skogan, 2021), making moral and social norms more salient (Senci et al., 2019; Hoeft et al., 2022; Klitgaard, 1988) and invoking a host of factors ranging from social image (Andreoni and Bernheim, 2009; Lacetera and Macis, 2010), and reputation concerns (Piazza and Bering, 2008a; Wu et al., 2016; Milinski et al., 2002) to guilt aversion (Falco et al., 2020). While transparency's effect via these channels is typically only assumed (Joshi, 2013), observability has been shown to reduce bribery (Garcia-Gallego et al., 2020), increase prosocial choices (Filiz-Ozbay and Ozbay, 2014; Andersson et al., 2020; Mill and Theelen, 2019; Kurzban and DeScioli, 2008), and increase altruistic punishment (Kurzban et al., 2007; Piazza and Bering, 2008b) even absent accountability. Yet these positive effects only hold if the observers are third parties (Bradley et al., 2018) while in our design, it is the community of the punisher acting as an observer.

Transparency may also backfire without the capability to punish those in power (Malesky et al., 2012; Hanna et al., 2010), potentially leading to resignation (Bauhr and Grimes, 2014) or less trust (Grimmelikhuijsen et al., 2013; Kolstad and Wiig, 2009). Such temporary increase in corruption measures has been observed after the enactment of freedom of information laws (Vadlamannati and Cooray, 2017), indicating that tolerating some hypocrisy might be beneficial (Bodnar and Salathhè, 2013). One example of how transparency can backfire is given by a study close to ours: In a public goods game absent peer punishment, an official embezzled more, if not just his identity, but the actions of all groups were common knowledge (Khadjavi et al., 2017).

Our study, therefore, contributes to the existing literature by filling a twofold gap. On the one hand, we deviate from the literature by looking at more subtle forms of harm through institutionalized power. Even if those in power do not abuse it by directly harming others, they can exempt themselves from what they enforce among the powerless. This kind of moral hypocrisy is both harder to detect and easier to justify. Even when one person has more power than the other, undercontribution and unequal earnings are seen as inappropriate over a large range of games (Cubitt et al., 2011; Kimbrough and Vostroknutov, 2016). Hoeft et al. (2019) further show that even punishers in our specific setting consider punishing contributions above their own as unacceptable. On the other hand, we take a first step in disentangling the mechanisms through which transparency can affect the behavior of the powerful. Those without power have no option to hold the punisher to account, as they can not react to his actions or punish him specifically.

3. Materials and methods

The goal of this paper is to investigate whether people in positions of power exploit it to exempt themselves from the duties they enforce onto others, which we will call *abuse of power*. Therefore, we need a setting with power asymmetries. Power is typically defined as asymmetric control over valued resources in a social relationship (Galinsky et al., 2015). To that end, we modify the established design of public-goods games with punishment to feature only one member with punishment power. As a novel feature of our design, this member was fixed for the entire experiment.[2]

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, next clarify the additional measurements, then discuss the predictions, 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 which they remained in for the duration of the public-goods game (partner-matching).[3] Thus, each group consisted of one punisher (labeled as D) and three non-punishers (labeled as A, B, and C), all in their fixed roles. The public-goods game was repeated for thirty rounds. Participants were instructed that each round would consist of three stages.

The first stage resembled a standard public-goods game. Participants were asked to allocate 20 tokens to a private and public account (1 token = 30 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") was informed about the first-stage contributions of all group members. In order to rule out reputation effects and potential targeted punishment based on previous behavior, the contributions of

[2] Note that we employ a rather implicit situation of power abuse, as the punisher is not communicated what exactly their duties are. In real-life situations, more explicit rules are given to people in power, and therefore officials and citizens know which actions can be considered abuse of power. We decided not to provide any guidance or expectations to punishers to keep the design clean. Specifically, telling the punishers what they are supposed to do would shift the behavior and would be subject to demand effects. Similarly, such a design choice would make a comparison to standard settings substantially more difficult.

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

the non-punishers were displayed to the punisher in random order each round.[4] D (the punisher) was now asked to indicate how much he would punish subject i (ς_i, $i \neq D$).[5] 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.[6] We set the low-power treatment to have enough punishment points to deter every participant from free-riding[7] to eliminate unobservable strategic considerations and focus on a purely behavioral effect. We set the high-power treatments such that the punisher can reduce the payoff of all non-punishers to zero in any scenario.[8] 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,[9] and more importantly, so the contributions of the punisher could be compared to the contributions of others directly.[10] 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 received 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.

3.1.2. Treatments

To study whether power corrupts and whether transparency might be able to reduce abusive behavior, we used a 2 (low power vs. high power) × 2 (low transparency vs. high transparency) between-subjects design.

We implemented the power treatments to answer whether power corrupts and absolute power corrupts more. In the baseline treatment, the punisher was equipped with 30 punishment tokens and thus, he was already powerful as he could make non-punishers indifferent between free-riding and fully contributing by punishing each with 10 tokens. To implement "absolute" power, the punishers were equipped with 120 punishment tokens, as 120 tokens were the upper bound of tokens ever needed to reduce the payoff of all non-punishers to zero. As both power treatments are strategically identical, the manipulation of power builds on a behavioral channel, as the punisher can destroy the other players vs. use just deterrent punishment. We denote the high-power treatments (where the punisher was equipped with 120 punishment tokens) by HighPwr and the low-power treatments (where the punisher was equipped with 30 punishment tokens) by LowPwr.

To answer the question of whether transparency can impede the corrupting effect of power, we implemented the transparency treatments. Transparency is typically defined as institutions ensuring that actions performed are easily observable. Accordingly, we vary the information available to those subject to power abuse. In the low-transparency treatments, participants were informed solely about the overall group contribution (in addition to their own contribution to the private and group account, their own punishment, and their payoff). Thus, non-punishers could merely see how much the group as a whole contributed and were not able to track the behavior of the punishers. Hence, punishers could hide behind the aggregated information and the resulting uncertainty for non-punishers. To manipulate transparency, participants were informed not of the overall group contribution but rather of the individual contributions of all members of the group. Thus, it was made transparent whether the punisher was abusing his power by contributing less than non-punishers. The high-transparency treatments (where all the contributions were public knowledge) will be denoted by HighTrans and the low-transparency treatments (where only the punishers knew individual contributions) by LowTrans.

[4] More specifically, the punisher only saw how much non-participants tagged "1", "2", and "3 contributed. Instead of labeling the non-punishers on this screen with their role assigned to them at the beginning of the experiment (A, B, or C), we tagged them "1", "2", and "3". For example, participant "A" was tagged "1" in round 4 and tagged "3" in round 5, etc. However, participants kept their internal labels throughout the experiment, and everybody (including the punisher) also observed the original labels (A, B, C, D) in the feedback stage if they were in the high-transparency treatment. Thus, relabeling non-punishers in this stage did not affect contribution dynamics and only ensured to rule out reputation-based punishment.

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

[6] The sum of all ς_i can at most be 30 in the "low" power treatment and 120 in the "high" power treatment.

[7] 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.

[8] Thus, this treatment presents the absolute power treatment as 120 tokens are the upper bound of tokens ever needed to reduce the payoff of all non-punishers to zero. This can be easily seen: if everybody (including the punisher) contributes fully, the payoff of every member would be 40. Thus, the maximum payoff the three non-punishers in combination can have is 120 tokens.

[9] 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. Furthermore, Kuwabara (2015) shows that average levels of punishment are almost identical with costly punishment and with costless punishment. Should we worry about costless punishment perhaps leading to more random punishment behavior? Kuwabara and Yu (2017) finds that patterns for costless punishment are similar to patterns observed in the literature with costly punishment. Also, in our experiment, we find patterns of punishment which are very similar to the patterns observed in Fehr and Gächter (2002), who used costly punishment (see Appendix A.3).

[10] Thus, we employ a costless punishment mechanism. The most important reason for doing so is that contributions of the punisher could be compared to the contributions of others directly. As noted in footnote 9, the punishment behavior in case of a costless and costly punishment mechanism seems to be comparable. Furthermore, we would bake in a mistake in our design if we were to employ a costly punishment mechanism, as the incentives of punishers and non-punishers would not be comparable anymore.

3.1.3. Additional measurements

We also collected data on spite (Marcus et al., 2014), rivalry & narcissism (Back et al., 2013), and social value orientation (SVO).

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 the slider-measure or the public-goods game task was chosen with equal probability to be payoff-relevant,[11] 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. Predictions

Our main goal in this paper is to investigate whether people in positions of power exploit their power to exempt themselves from the duties they enforce onto others, which we will call *abuse of power*. Abuse of power is the use of institutional power for illegitimate private gain. We define abuse as: *the deviation of the punisher's contribution from his own imposed contribution norm.*[12] Our definition mirrors the definition of legitimate punishment previously used, where punishers can only punish higher contributions (Faillo et al., 2013), and builds on the moral rejection of undercontribution (Cubitt et al., 2011; Kimbrough and Vostroknutov, 2016).[13] Hence, a punisher who imposes a norm of 18 in the current round, but contributes only 5, is behaving abusively. The amount of abuse is described by the difference between his imposed norm in the current round and his contribution in the respective round. In this example, it would be 13.

The imposed norm is the minimum contribution norm a punisher enforces. We consider punished contributions as violating the imposed norm, and unpunished contributions to fulfill the imposed norm.[14] For example, if a contribution of 18 is punished and a contribution of 19 is not punished anymore, the imposed norm is 18. Hence, we define the highest contribution still punished as the lower bound of the contribution norm.[15] 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.[16,17]

Rational choice theory, assuming selfish agents, predicts abusive behavior as punishment is costless.[18] Punishers will use their power to extort their peers. Apart from monetary incentives, punishers may have social or moral preferences. Conditional cooperation (Fischbacher et al., 2001; Fischbacher and Gachter, 2010; Herrmann and Thöni, 2009; Kocher et al., 2008), inequality aversion and social or moral norms (Cubitt et al., 2011; Kimbrough and Vostroknutov, 2016) constrain undercontribution by the punisher. For our purposes, we can integrate them in one moral cost term moderated by parameter m_i^c, the personal preferences of punisher i for avoiding the moral cost of undercontribution. This parameter also captures moral costs resulting from illegitimate punishment of higher contributions (Faillo et al., 2013). Essentially, the term reflects the moral costs of breaking the norm of undercontribution and illegitimate punishment.

Thus, for now, the punisher maximizes utility:

$$u_i(c_i, c_{i\neq j}, C^p, m_i^c) = (1 - m_i^c) \cdot (20 - c_i + 0.5 \cdot \sum_{j\in\{1,n\}} c_j) + m_i^c \cdot u^{Abuse}(c_i, c_{i\neq j}, C^p),$$

with c_i reflecting the punishers contribution, $c_{i\neq j}$ denoting the contribution of the other players, and C^p denoting the highest contribution still punished. We assume that $u^{Abuse}(\cdot)$, which denotes the disutility from breaking the norm of undercontribution and

[11] Hence, only one random problem was selected to become payoff-relevant following the arguments of Charness et al. (2016) and Azrieli et al. (2018). The argument for paying one random problem is threefold: 1) it is the only incentive-compatible mechanism (see Azrieli et al., 2018, for a detailed argument), 2) it reduces hedging, and 3) it increases the importance of each individual decision.

[12] In Appendix A.2, we also consider an alternative definition 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.

[13] Although shown in different settings, one may question whether participants share this opinion in our design. In a subsequent paper (Hoeft et al., 2019), we study how power abuse is perceived by subjects taking part in a similar game and subjects just learning the rules of the game. We show that – across all roles in the experiment – the social norm considers both under-contribution itself as well as punishment while undercontributing inappropriate.

[14] 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.3. For a detailed analysis of punishment norms, we refer the reader to Appendix A.4.

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

[16] Note that the results do not change by using a more lenient approach, in which a once established norm automatically continues to exist. This is sensible, since punishers create a threat-level that does not rely on continuous enforcement. Appendix A.5 shows further evidence of the effect of punishment on behavior in subsequent rounds. Appendix A.6 provides compelling evidence that punishers are very consistent in their punishment behavior.

[17] Note that a critical reader might point out that an alternative motivation might seem as abusive without it being the intention. Specifically, imagine that a punisher is a conditional cooperator and wants to ensure full contribution. In this case, in a given round, it might be that the punisher contributes to the same extent as the non-punishers in the previous round and at the same time punishes non-punishers in the current round as they have not reached the full contribution yet. This behavior would indeed be classified as abusive in our definition and might counter the intuitive understanding of abuse of power. There are two replies to this comment. First, such a motivation would indeed be classified as abusive for the first couple of rounds. However, as soon as the targeted contribution level is reached, the conditional cooperator would contribute to the same extent as the non-punishers and, thus, would not be classified as abusive anymore. Therefore, this concern vanishes over rounds. Second, we can empirically use an alternative definition of power abuse, in which a behavior is classified as abusive if the imposed norm in the current round exceeds the punisher's contribution in the subsequent round. Using this alternative definition does not change any of our results.

[18] Specifically, punishers are indifferent whether to punish or not in the last round. However, they can credibly signal that punishment will be used if full contribution is not implemented in the previous rounds. See also Hoeft et al. (2019) for a more formal argument.

illegitimate punishment, is decreasing in c_i, increasing in the difference between c_i and $c_{i\neq j}$, and increasing in the difference between c_i and C^p. The extent to which this disutility affects the overall utility of the punisher is denoted by the moral cost parameter m_i^c.

As the vast majority of participants care about moral concerns imperfectly (i.e., $m_i^c < 1$), we should still expect undercontribution and hypocritical punishment on average.

P0 *Punishers abuse their power, i.e., we predict that punishers undercontribute and punish hypocritically.*

In our low-power treatment, the punisher has just enough power to make everybody indifferent between defection and full cooperation. In our high-power treatment, the punisher can strip the participants of all their earnings in any given case.[19] Giving participants power has been shown to increase both selfish behavior and moral hypocrisy (see section 2). We, therefore, expect that those with higher power contribute less and engage in more hypocritical punishment. We can model this effect by adding a term that reduces the punishers' concern with moral costs $m_i^p \in (0,1)$, with m_i^p decreasing in power. Our term can be understood as the disinhibition of immoral desires and an increased focus on the self caused by power (Lammers et al., 2015).

Thus, the punisher maximizes utility

$$u_i(c_i, c_{i\neq j}, C^p, m_i^c, m_i^p) = (1 - m_i^c \cdot m_i^p) \cdot (20 - c_i + 0.5 \cdot \sum_{j\in\{1,n\}} c_j) + m_i^c \cdot m_i^p \cdot u^{Abuse}(c_i, c_{i\neq j}, C^p)$$

Hence, no power would result in a power concern of $m_i^p \approx 1$, and hence, the punisher would care fully about the moral costs; Having infinite power would result in $m_i^p \approx 0$, i.e., the punisher ignoring moral costs. This is in line with abusive behavior scaling with power in the sense of "Power corrupts. Absolute power corrupts absolutely." (John Dalberg-Acton).

An alternative interpretation of the effect of power would follow some of the literature indicating that some power corrupts, but a lot of power makes people more responsible (Sassenberg et al., 2014, 2012; Tost et al., 2015). In particular, it follows the observation that participants act selfishly in an ultimatum game up to the point where the recipient is completely powerless (Handgraaf et al., 2008). Thus, we would expect m_i^p to have a convex form with no power resulting in a punishers' concern with moral costs of $m_i^p \approx 1$, some power resulting in a concern of moral costs of $1 > m_i^p \geq 0$, and infinite power making people more responsible and moving $m_i^p \to 1$.

Thus, we have two competing predictions on the effect of power. If the punishers' concern with moral costs decreases linearly in power, we expect the high-power treatments to increase abusive behavior always. Alternatively, if the punishers' concern with moral costs changes in a convex way, we might predict high-power treatments to increase abusive behavior, but we also might expect high-power treatments to decrease abusive behavior if the punisher is too powerful.

P1a *Higher power leads to more abuse, i.e., we predict more undercontribution and a higher imposed norm in the high-power treatments.*
P1b *Higher power leads to less abuse, i.e., we predict less undercontribution and a smaller imposed norm in the high-power treatments.*

Beyond the effect of power, we test whether transparency can have an effect without accountability. In our design, transparency enables participants to see the individual contributions to the public good and thereby detect unfair contribution patterns. Yet they cannot act on this knowledge as there are no institutions, third parties or options for peer punishment to hold the punisher accountable.[20] If we find a positive effect of transparency, it is bound to be due to psychological mechanisms mentioned in the literature section (e.g., social or self-image concerns, reputation, guilt aversion, shame, etc.).

We do not aim to distinguish between the different proposed psychological mechanisms in this study for different reasons: Enabling accountability is generally seen as the primary function of transparency and it remains in doubt that it can change anything in its absence. Establishing the existence of other pathways is a necessary first step. Previous research has shown that audience effects do matter, but only in settings without punishment power (Filiz-Ozbay and Ozbay, 2014; Andersson et al., 2020; Kurzban and DeScioli, 2008) or for third parties who have no opportunity for hypocritical punishment (Kurzban et al., 2007; Piazza and Bering, 2008b). As referenced in the literature section, more than a handful of different psychological mechanisms have been suggested. Distinguishing among them does not seem feasible in our design, which is already fairly complex and long. Even simple additional elements, such as belief elicitation, may prime social or moral norms, activate guild aversion, or distract from abusive behavior that would otherwise be noticed. As is custom in the literature on transparency, we rather test transparency as an institution. As such, transparency always changes more than one variable at once.

For the purpose of our model, the effect of transparency does not depend on the specific psychological mechanisms. Since contribution as much or more than others does not violate any social preferences, moral or social norms, it does not matter whether social or self-image, reputation, guilt aversion, shame, etc. drive the result. In any case, it should increase the weight punishers place

[19] We decided to give the low-power punisher enough punishment points to deter everybody to eliminate unobservable strategic considerations. Still, in our high-power treatments the punisher can enforce his will beyond mere indifference. However, we would expect the same result otherwise as simply priming higher power is enough to cause its detrimental effects (Galinsky et al., 2003).

[20] Non-punishers may try to resist by withholding contributions. As this could be a reaction to the behavior of any PGG member and would impact everybody, it hardly constitutes accountability. And as our punisher has enough (costless) power to make all indifferent between contribution and non-contribution, and coordination is impossible, any such effect would have to be psychological as well.

on avoiding moral cost, which we can capture by adding a term m_i^t to the moral cost parameter. In this way, our model is also consistent with others that posit an interaction between transparency and a moral cost term (Falco et al., 2020; Garcia-Gallego et al., 2020; Salmon and Serra, 2017).

Thus, the punisher maximizes utility

$$u_i(c_i, c_{i\neq j}, C^p, m_i^c, m_i^p, m_i^t) = (1 - (m_i^c \cdot m_i^t) \cdot m_i^p) \cdot (20 - c_i + 0.5 \cdot \sum_{j\in\{1,n\}} c_j) + (m_i^c \cdot m_i^t) \cdot m_i^p \cdot u^{Abuse}(c_i, c_{i\neq j}, C^p)$$

We assume $m_i^t \in [0,1]$ to be the weight put on moral costs due to transparency, and to be decreasing in transparency. Hence, under absolute hidden behavior (i.e., nobody will ever know what happened), we assume $m_i^t = 0$, and the behavior of punishers is driven by their self-interest. With increasing transparency, the weight on the overall moral costs (due to self-image, reputation, guilt aversion, shame, etc.) increases. Thus, in line with the majority of research on transparency, we expect transparency to increase the moral or social costs of antisocial behavior.[21]

P2 *Transparency leads to less abuse, i.e., we predict less undercontribution and a lower imposed norm in the high-transparency treatments.*

Focusing on the combination of power and transparency, it is straightforward how higher power can impede transparency in our model. Since power makes people care less about the moral cost of their action, it will equally impede the effect of transparency. This is captured by the intuition that transparency might induce shame, but will not deter the truly shameless (Fox, 2007). A similar logic applies to other social or moral concerns. As discussed above, those in power fail to live up to their moral standards, employ hypocritical evaluations and care less about others and their opinions.

In terms of the model, we have two counteracting effects. Transparency increases the weight put on the moral costs, i.e., $\frac{\delta m_i^t}{\delta t} > 0$, and thus deters undercontribution. Power, on the other hand, might reduce the weight put on the moral costs if the concern for moral costs is monotonically falling in power ($(m_i^p)' < 0 \wedge (m_i^p)'' \leq 0$), but it also might increase concerns for moral costs if the effect is convex ($(m_i^p)'' > 0$). Thus, power might induce undercontribution.

In the first case, we would expect power to impede the effect of transparency. In the second case, we would also expect power to impede the effect of transparency if we are below the "threshold" (i.e., $(m_i^p(x_0))' < 0$). However, if we are above the "threshold" (i.e., $(m_i^p(x_0))' > 0$), then we would expect power to deter undercontribution and increase the effect of transparency. This can be thought of as transparency leading participants to understand their position of power as one of responsibility (Sassenberg et al., 2014, 2012; Tost et al., 2015). However, it is unclear when the threshold is reached, which makes us predict power to impede the effect of transparency.

P3 *Higher power impedes the effect of transparency, i.e., we predict the positive effect of transparency on punishers' contributions to be weaker under high power.*

Importantly, we have so far assumed that the effect of power is not directly a function of transparency. However, making power discrepancies transparent may make high power especially salient. Thus, the convex function of power might itself be a function of transparency, lowering the threshold point after which power is not considered an opportunity but rather a responsibility. If that were to be the case, we might observe a small effect of transparency under low power and a high positive effect of transparency under high power.

3.3. 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 seating, 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.[22] 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.

[21] While we discussed some literature showing how transparency may backfire, this concerned mostly the reactions of the powerless. Our punishers also have no objective reason to feel threatened, even in full transparency. While (Khadjavi et al., 2017) did find that transparency increases embezzlement, their design differs significantly from ours. On the one hand, their transparency treatment changes the information of the official as well. All observed changes could therefore be driven by the official learning new information about the powerless, instead of vice versa. Furthermore, the official may embezzle in an attempt to punish other group members for low contributions. He may do this even where he could peer punish, as embezzlement is costless for him while punishment is not.

[22] 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.

On average subjects had a SVO-score of 15 (on a scale from -16.26 (fully competitive) to 61.39 (fully altruistic)). Further, subjects had a spite-score of 2 (on a scale from 1 to 5) and a rivalry-score of 21 (on a scale from 9 to 54) on average. The average narcism-score was 49 (on a scale from 18 to 108). All subjects' characteristics are also shown and compared separately for all four treatments in section A.1.

4. Data analysis

The subsequent data analysis will be structured as follows: We will start by studying abuse in the four treatments. Thereafter, we will investigate the imposed norm of punishers and finally, we will take a look at the contribution of punishers and non-punishers in the four treatments.[23]

Our data will demonstrate high levels of abuse overall, with high power having a corrupting effect under low transparency. Combining high 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 they themselves undercut their norm and consequently the contributions of non-punishers.

Econometric model To highlight treatment differences, we will present simple mean comparisons, and we will also use a more structured estimation model to obtain further insights. More specifically, we will be able to move beyond average treatment effects and study how the treatments differ over time, accounting for the nested structure of the data. The following shows a formal description of the econometric model (see also Snijders and Bosker, 2011, p. 69):

$$\begin{aligned} Y_{i,k,t} =& \beta_0 + \beta_1 \cdot \mathbb{1}_{\text{HighTrans}} + \beta_2 \cdot \mathbb{1}_{\text{HighPwr}} + \beta_3 \cdot \mathbb{1}_{\text{HighTrans}} \cdot \mathbb{1}_{\text{HighPwr}} \\ &+ \beta_4 \cdot t + \beta_5 \cdot t \cdot \mathbb{1}_{\text{HighTrans}} + \beta_6 \cdot t \cdot \mathbb{1}_{\text{HighPwr}} + \beta_7 \cdot t \cdot \mathbb{1}_{\text{HighTrans}} \cdot \mathbb{1}_{\text{HighPwr}} \\ &+ \beta_7 \cdot \mathbb{1}_{t=30} + \epsilon_k + \epsilon_{i,k} + \epsilon_{i,k,t} \end{aligned} \tag{3}$$

$Y_{i,t}$ represents the behavior of subjects i in round t in group k with $i \in \{1, \dots, n\}$, $t \in \{0, \dots, 30\}$. $\mathbb{1}_{\text{HighPwr}}$ denotes a dummy with value one if the participants are in the high-power treatment, i.e., punisher is equipped with 120 punishment tokens, and zero if the participants are in the low-power treatment, i.e., punisher is equipped with 30 punishment tokens. $\mathbb{1}_{\text{HighTrans}}$ denotes a dummy with value one if the treatment is with high transparency and zero if the treatment is low transparency. t denotes the period effect, with $t \in \{0, \dots, 30\}$. $\mathbb{1}_{t=30}$ indicates a last-round control, as the contribution and abuse behavior might differ in the last round. $\epsilon_{i,k,t}$ denotes the residuals and follows the typical assumption that it is normally distributed with mean zero ($\epsilon_{i,k,t} \sim \mathcal{N}(0, \sigma_3^2)$).

The econometric model shown in Equation (3) is a mixed-effects model (also sometimes called random-effects or multilevel model), which is a common approach in a situation where it is essential to account for the nested structure of the data. One particularly useful feature of the mixed-effects model is that it can account for the fact that every participant might have a different intercept (e.g., due to individual levels of prosocial preferences) and every group might have a different intercept (e.g., due to the random composition of the group). By using a mixed-effects model with a random intercept per group and a random intercept per individual within that group, we obtain estimates with standard errors clustered on the group level while still making use of the individual-level data. These two random effects are also represented in the econometric model of Equation (3). Specifically, ϵ_k denotes the group's random-intercept effect accounting for each group's idiosyncratic changes (which might be driven by idiosyncratic punishers' behavior). ϵ_i denotes the random-intercept effect of each individual within a group accounting for every participant's idiosyncratic intercept.

A similar approach could have been running an OLS with cluster-robust standard errors at the group level.[24] However, such an approach is typically less efficient. In fact, the econometric literature suggests the use of a mixed-effects model over a simple clustering approach (Fitzmaurice and Laird, 2014; Gelman and Hill, 2013). For example, Bell et al. (2018) and Heisig et al. (2017) show that theoretically and empirically mixed-effects models are superior to fixed-effects approaches. Heisig et al. (2017), for example, write: "OLS with cluster-robust standard errors delivered the poorest performance, in terms of both precision and inference. [...] A flexible mixed effects specification [...] produced the most precise estimates". Therefore, we use a mixed-effects model with a random group intercept and random individual-level intercept within that group throughout our paper. In cases where we focus on the punishers' behavior, we only use a group-level random intercept model, as there is only one punisher per group.

Note that one assumption made in the model is that the time trend is linear. We will see that this assumption seems sufficiently reasonable. However, we also relax this assumption in Appendix B. For that purpose, we estimate a common loess spline for the dependent variable over rounds over all treatments using a Bayesian approach (see Kirchkamp and Mill, 2020, 2021, for a similar approach). All results are basically identical to the results reported in a linear trends non-Bayesian approach.

4.1. Abuse of punishment power

Our main goal in this paper is to investigate whether punishers abuse their position of power and what determinants affect this abuse. As a reminder: We define abuse as the deviation of the punisher's contribution from his own imposed contribution norm.

[23] If not mentioned otherwise, we will use two-sided t-tests on variables aggregated over all rounds by groups. The results remain stable using non-parametric tests.

[24] In our case, this approach leads to the same estimates but typically higher standard errors.

Table 1
Descriptives of abuse.

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

Note: The table displays the fraction of punishers ever punishing contribution above their own contribution (i.e. abusing) in columns two. Column three denotes the average frequency of abuse across all rounds and punishers who abused their power at least once.

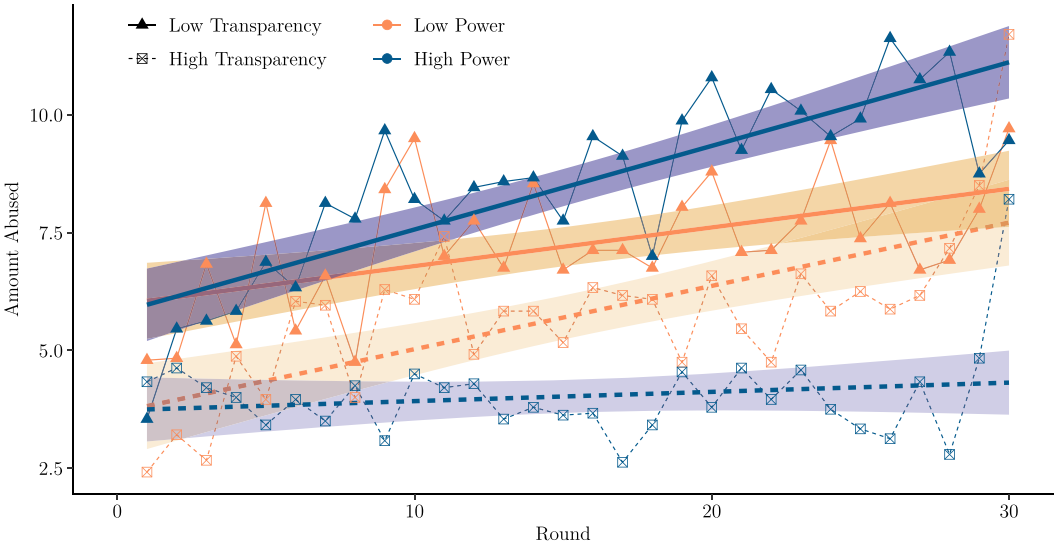

Fig. 1. Amount abused in each of the four treatments over time. Blue lines represent the punisher's abusive behavior in the high-power treatments, while orange lines represent the low-power treatments. The high-transparency treatments are represented by dashed lines with crossed cubes, while the low-transparency 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. The corresponding tunnels surrounding the linear regression lines represent the 95% confidence intervals. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Thus, the difference between the contribution of punisher and the highest contribution of non-punishers still punished (the imposed contribution norm) represents the amount of abuse. Hence, a punisher who punishes a contribution of 18 (thus imposes a norm of 18), but contributes only 5, is behaving abusively to the amount of 13.

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, $t(46) = 2.5$, $p = 0.015$, $d = 0.7$ (medium); however, it has no significant effect on the average frequency of abusive behavior under low power $t(46) = 0.8$, $p = 0.424$, $d = 0.2$ (small). Under low transparency power, on average, also seems to not significantly influence abuse ($t(46) = 1.0$, $p = 0.317$, $d = 0.3$ (small)). However, as can be seen in Fig. 1, these differences in abusive behavior are visible and are increasing over time.

So far we have only seen how many subjects decided to abuse, but we have not seen the amount of abuse. As can be seen in Fig. 1 there seems to be no level effect of transparency in the low-power treatment on amount abused ($t(46) = 0.9$, $p = 0.39$, $d = 0.2$ (small)) while there is a level effect of transparency in the high-power treatment ($t(46) = 2.7$, $p = 0.011$, $d = 0.8$ (medium)). On average, the amount abused is lowest in the high-power high-transparency treatment – with $M = 4.03$ ($SD = 5.45$) points of abuse – and the highest in the high-power low-transparency treatment – with $M = 8.54$ ($SD = 6.29$) points of abuse. Both low-power treatments are in between these two extremes (low-power low-transparency: $M = 7.24$ ($SD = 5.87$); low-power high-transparency: $M = 5.76$ ($SD = 5.92$)).[25]

[25] Note that our results are not driven by very few individuals who behave "crazy". In fact, we see that the majority of groups behave abusively more than 10, 15, and even 20 of the 30 rounds. The treatment differences might be driven by participants changing their type. Specifically, we see that 71%, 79% of participants are behaving abusively in more than 10 rounds in the low-power low-transparency treatment and high-power low-transparency treatment, respectively. We also see that 54%, 42% of participants are behaving abusively in more than 10 rounds in the low-power high-transparency treatment and high-power high-transparency treatment, respectively. Thus, abusive behavior is very predominant, and our results are not driven by a small number of groups behaving oddly.

Table 2
Mixed-effects model of the abusive behavior.

	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* (0.05)	0.17*** (0.02)	0.06*** (0.02)
t x HighTrans	0.05* (0.03)	0.06** (0.03)	−0.16*** (0.02)	0.05* (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	×	✓	×	×
Group specific Effects	✓	✓	✓	✓
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: $^{*}p < 0.1;^{**}p < 0.05;^{***}p < 0.01$;
Controls include age, gender, spite, SVO, narcissism, rivalry, and the interaction of those variables with the period. HighTrans denotes a dummy with value one if the treatment is with high transparency and zero if the treatment is low transparency. HighPwr denotes a dummy with value one if the participants are in the high-power treatment, i.e. punisher is equipped with 120 punishment tokens, and zero if the participants are in the low-power treatment, i.e. punisher is equipped with 30 punishment tokens. *LastRound* indicates a last-round control. *t* denotes the period effect. The first two columns show the abuse behavior in all four treatments, while the third and fourth column shows the abuse behavior in the high and low power treatments, respectively. To account for the repeated nature of the interactions, we use a mixed-effect model following Equation (3), with a group-specific random effect.

To estimate the time trends, we use the linear mixed-effects model from Equation (3) with fixed effects on treatments, time, the interaction of time and treatments, and a control for the last round, as well as group-specific random effects. The results are 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. The non-linear Bayesian approach, reported in Appendix B.3.1, further supports the robustness of the estimates.

We can see that abusive behavior increases over time. Hence, subjects learn to abuse their power in all treatments except for the high-power high-transparency treatment.

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

Surprisingly, the effect of transparency was the opposite of our expectations (prediction 2): under low power, it marginally increased abusive behavior. Hence, transparency was not only not helpful, it was actually harmful in the low-power treatment. This finding might be rationalized by those with low power reacting badly to any constraints on their power as they perceive their position to be unstable (Foulk et al., 2020; Fast et al., 2012; Williams, 2014; Maner and Mead, 2010).

Concerning our third prediction, 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 1a 83.3% of all punishers abused their power at least once and abuse increased over time.
Result 1b High transparency has a marginally significant positive effect on abusive behavior in the low power setting.
Result 1c The increase in abusive behavior over time is stronger under high power compared to low power.
Result 1d Increased transparency reduces abusive behavior significantly in the high power setting.

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

4.2. Imposed norm

In this section, we investigate the norms punishers enforced.[26]

We measured for all punishers the norm they imposed. As a reminder, the imposed norm is the cutoff level of contributions below which members are punished, i.e., the highest contribution punished. The average development over time of those norms by

[26] 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.3. For an analysis of the aggregate punishment norm in each treatment, see Appendix A.4.

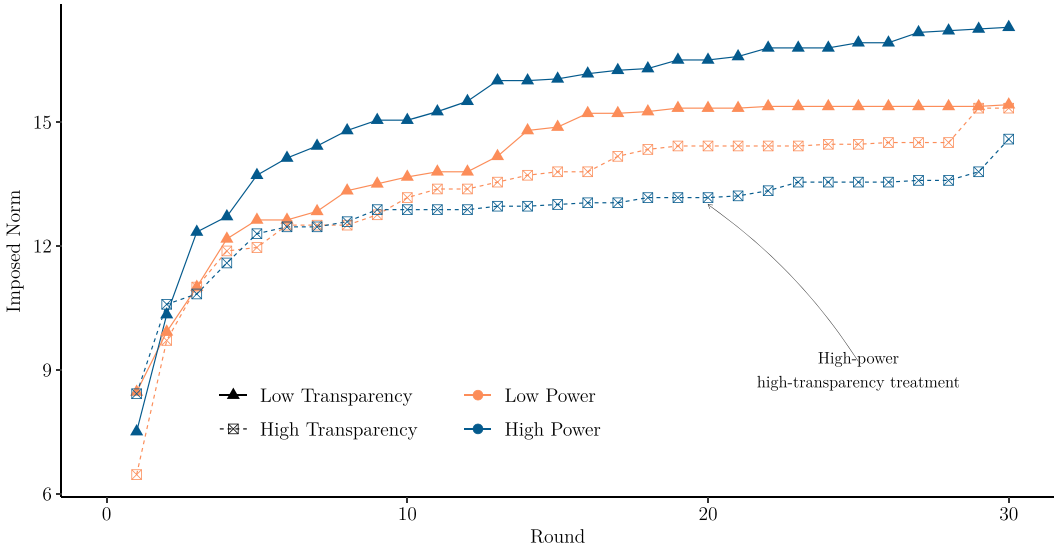

Fig. 2. Imposed norms over time in each of the four treatments. Blue lines represent the punisher's imposed norms in the high-power treatments, while orange lines represent the low-power treatments. The high-transparency treatments are represented by dashed lines with crossed cubes, while the low-transparency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

treatment is shown in Fig. 2. By definition, the norm either stabilizes or increases. After roughly 5 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. Appendix A.6 discusses the consistency of punishers in detail and shows that punishers are very consistent in their punishment behavior. Specifically, we find that punishers who punish a certain contribution in a given round are very likely to punish this contribution – or any other contribution lower than this contribution – in the subsequent period. In fact, there is a 90% probability of receiving punishment for a contribution that was previously punished. Moreover, we look at the dynamics of punishment and can demonstrate that punishers primarily focus their punishment on the lowest contributors and all non-punishers who contribute below the imposed norm – the punishers only rarely punish all non-punishers. All of this evidence speaks in favor of a very consistent behavior by punishers: punishers steadily increase their imposed norm and very consistently punish any violations by non-punishers by precisely striking those who contribute below the imposed norm.

To estimate the effect of the treatments on the imposed norm, we estimate a mixed-effects model following Equation (3). We again estimate a linear regression for the imposed norm with fixed effects of treatments, time, the interaction of both and controlling for group-specific random effects. The results are reported in Table 3. Further, we also estimate the effects using a non-linear Bayesian approach reported in Appendix A.7. We see that the average imposed norm in the first round is slightly higher than half of the endowment. Over time, this norm increases. We also see 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.

In Table 3, we concentrate mainly on the trends. As can be seen in Fig. 2 there seems to be no level effect of transparency in the low-power treatment on the imposed norm, $t(46) = 0.5$, $p = 0.608$, $d = 0.1$ (negligible) while there is a rather obvious level effect of transparency in the high-power treatment, $t(46) = 2.1$, $p = 0.044$, $d = 0.6$ (medium). Again, on average, the norm imposed is lowest in the high-power high-transparency treatment – with $M = 12.78$ ($SD = 5.01$) points – and the highest in the high-power low-transparency treatment – with $M = 15.34$ ($SD = 3.40$) points. Both low-power treatments are in between these two extremes (low-power low-transparency: $M = 14.05$ ($SD = 4.63$); low-power high-transparency: $M = 13.32$ ($SD = 5.17$)).

Overall, the imposed norm (i.e., the cutoff level of contributions below which members are punished), seems to be mostly established within the first five rounds and then slowly increases. The median imposed norm across all conditions and rounds equates to 15, which means that contributions of 75% of the endowment are still punished. It is informative to contrast this norm to the literature. The vast majority of the literature studies norms concerning the punishment of freeriders (Cubitt et al., 2011), or the deviation from the own contribution (Hauge, 2015) or the mean contribution (Carpenter and Matthews, 2012). Punishing above the own contribution (defined as abusive behavior in our setting) is typically considered antisocial (Zhou et al., 2017; Albrecht et al., 2018) or hypocritical punishment (Carpenter, 2007; Burton-Chellew and Guérin, 2021; Mieth et al., 2021). Also, most participants prefer to have no punishment institution or an institution punishing only the lowest contributors (Ertan et al., 2009). Some literature studies what is considered an appropriate contribution and find that small contributions are clearly considered inappropriate (even if contributing is inefficient Abbink et al., 2017). Kimbrough and Vostroknutov (2016) have results showing that the consideration of appropriate behavior is increasing in contribution. However, contributing more than half of the endowment is considered appropriate. Also Hoeft et al. (2019) show that punishing contributions below the own, in our setting, is considered socially inappropriate across subjects. This stands in contrast to our findings as punishers, on average, still punish contributions of 75% of the endowment – and they do so antisocially. Similarly, Otten et al. (2020) measure what contribution is considered appropriate in heterogeneous games,

Table 3
Linear mixed-effects model of the imposed norm.

	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	×	✓
Group specific Effects	✓	✓
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: $^{*}p < 0.1;^{**}p < 0.05;^{***}p < 0.01$;
Controls include age, gender, spite, SVO, narcissism, rivalry, and the interaction of those variables with the period. HighTrans denotes a dummy with value one if the treatment is with high transparency and zero if the treatment is low transparency. HighPwr denotes a dummy with value one if the participants are in the high-power treatment, i.e. punisher is equipped with 120 punishment tokens, and zero if the participants are in the low-power treatment, i.e. punisher is equipped with 30 punishment tokens. To account for the repeated nature of the interactions, we use a mixed-effect model following Equation (3), with a group-specific random effect.

Table 4
Fraction of punishers imposing specific norms.

Treatment	Imposed norm ≤ 5	Imposed norm ≥ 15	Imposed norm ≥ 19	Groups
LowPwr x LowTrans	4	75	33	24
LowPwr x HighTrans	4	62	38	24
HighPwr x LowTrans	0	88	46	24
HighPwr x HighTrans	8	67	17	24

Note: The table displays the fraction of punishers punishing a contribution below or equal to 5, or above or equal to 15 and 19 in the last 5 rounds of the experiment by treatment.

and find that equal contribution or equal earnings are considered appropriate. Translating this insight to our setting, the punisher would be expected to match the contribution of others – which, again, clearly is not the case. Finally, we can compare our induced norm to estimated punishment norms in Carpenter and Matthews (2009). Carpenter and Matthews (2009) find that the best predicting punishment norm is to punish contributions below 75% of the endowment. This number is strikingly similar to our median imposed norm in our setting. However, we also estimate the punishment norm in the way as Carpenter and Matthews (2009) in Appendix A.4, and also find that an absolute norm is best-describing punishment norm. However, in our estimate, punishing everything short of full contribution is the best predicting norm, indicating that our punishers expect more contribution than typically found in benevolent punishment behavior in PG games.

Although the average of the imposed norm across all punishers provides a general overview, it also obscures some of the variance. Table 4 outlines the proportion of punishers instating a norm of 5 or under, or a norm of 15/19 or above during the final five rounds within each treatment. The majority of punishers in all treatments impose a contribution greater than 15. More strikingly, almost half of all punishers in the high-power low-transparency treatment expect full contribution (meaning, a contribution of 19 is still subject to punishment). These patterns align with the results above, as the fewest punishers enforcing near-total contribution are found in the high-power high-transparency treatment, and the two low-power treatments falling again in between. However, we also observe heterogeneity in the results, as there are punishers who endorse only a very minimal contribution norm (5 or below). These results again neatly reflect treatment differences, with the smallest number of punishers enforcing a minimal contribution in the high-power low-transparency treatment. This is then followed by the two low-power treatments, and, subsequently, the high-power high-transparency treatment.

In Appendix A.7, we take an even closer look at the heterogeneity in the imposed norm. While Fig. 2 seems to suggest that the imposed norm is consistent across rounds and groups, the evidence of the heterogeneity in the data (see also Figure 11 in the Appendix) speaks against an endogenously accepted norm across groups. Further, we also study how many non-punishers adhere to the norm in Figure 12 in the Appendix, and find a sizable portion of non-punishers accepting the norm. However, we also find considerable heterogeneity, with more than 30% of groups having at least one non-punisher violating the imposed norm. All of this suggests that punishers impose different norms, but the imposed norms follow the treatment conditions.

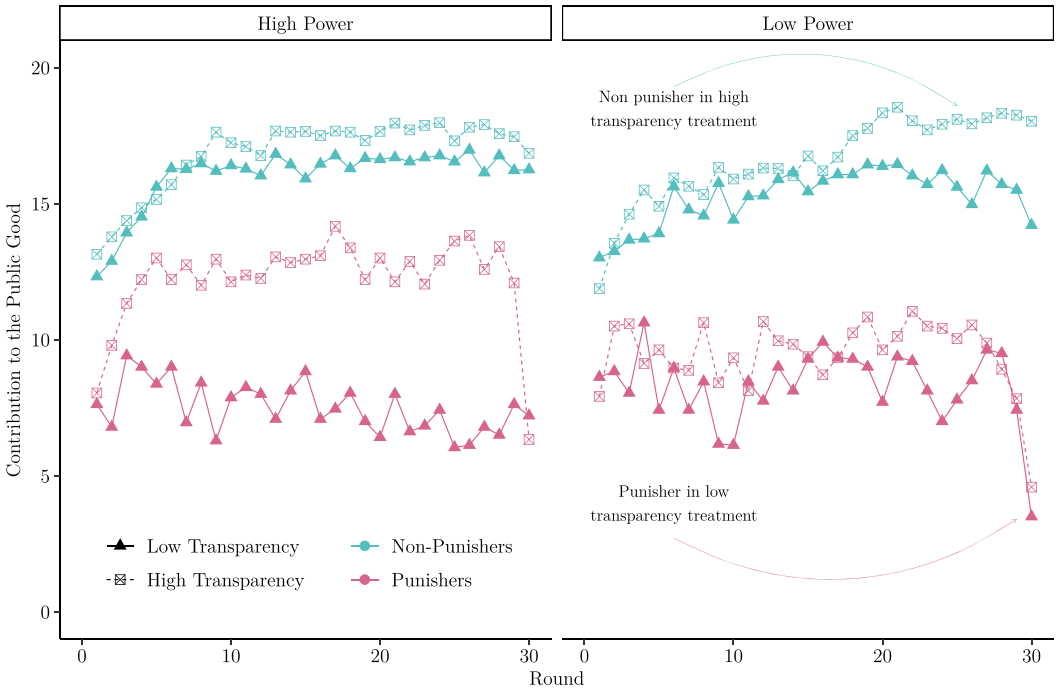

Fig. 3. Contribution to the public good over time in the respective treatments. Red lines represent the punisher's contributions, while sky blue lines represent the non-punisher's contributions. The graph on the right shows the contribution behavior in the low-power treatments, while the graph on the 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.

Result 2a The imposed norm increases over time.
Result 2b Increased transparency does not affect the imposed norm under low power.
Result 2c High power leads to higher imposed norms over time.
Result 2d High transparency reverses the effect of high power on the imposed norms.

Summarizing this section, it seems like the abusive behavior is driven by the norm punishers imposed. The imposed norm is highest 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

One open question is how the imposed norm affects contribution behavior, and how contribution behavior changes over time.

The punisher can influence the contribution behavior of non-punishers in essentially two ways. By acting as a role model and contributing a lot (leadership-effect), or acting as a punisher and forcing contribution behavior through penalties (punishment-effect). In Appendix A.5, we focus on these two channels in detail. We find only negligible leadership-effects, as the contribution behavior of non-punishers remained largely unaffected by the punishers' contribution in both low and high-transparency environments. There was only a minor increase when punishers contributed exceptionally high under the high transparency conditions.

More importantly, we find that punishment has a striking influence on non-punishers' contributions. Non-punishers substantially increase contributions following punishment and decrease them slightly in its absence. The severity of punishment directly correlates with the magnitude of increased contributions. These patterns persist across different treatments, indicating punishment as an effective motivator for increased contributions.

To see how contribution behavior changes, we will now focus on the contribution behavior of all participants. Fig. 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 ($\overline{SD}^{27}$ = 3.57) while non-punishers contributed M = 16.19 ($\overline{SD}$ = 4.28) points to the public good, a highly significant difference: $t(190) = 7.9$, $p \leq 0.001$, $d = 1.1$ (large).[28]

[27] Average of the standard deviations over all rounds.
[28] This effect has also been reported in Hoeft and Mill (2017) for the low-power treatments.

Table 5
Linear mixed-effects model of the contribution to the public good.

	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	×	✓	×	✓
Group specific effects	✓	✓	✓	✓
Sbj specific Effects	✓	✓	×	×
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;
Controls include age, gender, spite, SVO, narcissism, rivalry, and the interaction of those variables with the period. HighTrans denotes a dummy with value one if the treatment is with high transparency and zero if the treatment is low transparency. HighPwr denotes a dummy with value one if the participants are in the high-power treatment, i.e. punisher is equipped with 120 punishment tokens, and zero if the participants are in the low-power treatment, i.e. punisher is equipped with 30 punishment tokens. *LastRound* indicates a last-round control. *t* denotes the period effect. The first two columns show the contribution behavior of non-punishers in all four treatments, while the third and fourth column show contribution behavior of punishers. For non-punishers, subject specific fixed effects and group specific effects are controlled for. To account for the repeated nature of the interactions, we use a mixed-effect model following Equation (3), with a group-specific random effect.

Discriminating by treatments in Fig. 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. Concerning the average contribution over all rounds, we find the same effects. While the contribution is rather stable in all three treatments[29] it is significantly higher in the high-power high-transparency treatment (with $M = 12.25$ ($SD = 7.91$) points of contribution to the public good, $t(94) = 2.2$, $p = 0.03$, $d = 0.5$ (medium)). This finding might be rationalized by punishers only in the high-power treatment understanding their position of power as one of responsibility instead of opportunity (Sassenberg et al., 2014, 2012; Tost et al., 2015), which makes them contribute more.

Non-punishers start with similar contributions in all treatments and increase their contributions over time.[30] While high transparency generally strengthens the increase in contributions, the combination with high power dampens this positive effect significantly.[31]

To estimate the effect of the treatments on the contribution behavior over time, we estimate a mixed-effects model following Equation (3). The estimations are reported in Table 5. All results are robust to the inclusion of controls (age, gender, spite, rivalry, narcissism, and SVO, and the interaction of the mentioned factors with time). The results are also robust to using a non-linear Bayesian approach (see Appendix B.1.1 and B.1.2).

We observe that the behavior of punishers is very stable and does not differ between the treatments except in the high-power high-transparency treatment, where contribution behavior is increasing significantly over time. Non-punishers, on the other hand, increase their contribution significantly over time in all treatments. This increase is stronger in the low-power high-transparency treatments and significantly smaller in the high-power high-transparency treatment.

In Appendix A.8 further zoom in into the contribution behavior. We make three interesting observations. First, there is substantial heterogeneity in terms of contribution behavior. While most groups contribute more than 50% of their endowment, there are groups who contribute substantially less. Second, a substantial fraction of groups reach the socially optimal full contribution behavior.

[29] Low-power low-transparency: $M = 8.29$ ($SD = 6.97$); low-power high-transparency: $M = 9.48$ ($SD = 7.89$); High-power low-transparency: $M = 7.51$ ($SD = 6.85$)).

[30] Looking at the average contribution of non-punisher, we find that the contribution levels are very similar. In the high-power high-transparency treatment non-punishers contributed $M = 16.87$ ($SD = 3.76$) points while the high-power low-transparency treatment non-punishers contributed $M = 16.02$ ($SD = 4.05$) points to the public good. In the low-power high-transparency treatment non-punishers contributed $M = 16.62$ ($SD = 3.34$) points while the low-power low-transparency treatment non-punishers contributed $M = 15.27$ ($SD = 4.62$) points to the public good.

[31] 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. In Appendix A.5 we further investigate whether punishment is causal for increased contribution by non-punishers and find very strong evidence for this claim.

Across all treatments, more than 50% of groups reach full contribution. However, there are differences by treatments. In particular, we observe the highest full contribution levels in the high-power high-transparency treatment. Further, the contribution is highly divided between punishers and non-punishers, with a clear majority of non-punishers contributing fully and only a fraction of punishers doing the same. Third, there does not seem to be a clear understanding of punishers of how to behave. In particular, the contribution behavior of punishers is very polarized, with the two biggest camps being full contributors and full free riders. This also indicates that much of the treatment differences in punishers' contribution behavior are driven by the extensive margin (i.e., contributing or not). This all speaks against one underlying theoretical norm in this game that subjects agree upon. It is rather that contribution behavior is polarized, but the treatments change the types of punishers, leading to the highest fraction (50%) of full contributors in the high-power high-transparency treatment and the lowest (22%) in the high-power low-transparency treatment.

Result 3a Punishers contribute far less than non-punishers. Non-punishers increase their contributions over time, while punishers do not.

Result 3b Increased transparency leads to a stronger increase overall for non-punisher and does not affect the contribution behavior of punishers.

Result 3c Higher power does not significantly change the contribution behavior of punishers or non-punishers.

Result 3d High power decreases the positive effect of transparency on the contributions of non-punishers, while it increases the contributions of punishers over time only.

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 contributions stay virtually the same.

Taken all the results together, we see that 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 to enforce prosocial behavior. The power conferred by these institutions can be exploited for selfish gain. While outright extortion of others may be easy to monitor and avoid, those in power can enforce rules they do not adhere to themselves and thereby "play by different rules". Such hypocritical punishment (Lammers et al., 2010) constitutes abuse of power and compounds inequality as those already in power contribute less to public goods.[32] We provide a first investigation of institutional second-party punishers in a repeated game, and indeed find a large degree of power abuse. Punishers enforce high contribution norms, but only contribute half of what they expect from others.

In line with our predictions, we find that an 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 prediction: Instead of limiting abuse for low-power treatments, it did so only for the high power one. This is a promising result from a policy perspective, as transparency can work even when power discrepancies are high and true accountability unrealistic. Yet it suggests that the relationship between power and transparency is more complex than previously thought. While a large majority of previous findings in social psychology highlighted the corrupting effect of power, a few studies have indicated that there may be boundary conditions to this effect: If power is construed as a responsibility rather than an opportunity, it leads to more prosocial behavior (Sassenberg et al., 2014, 2012; Tost et al., 2015). This can be expected more from those with high power: Participants act selfishly in an ultimatum game up to the point where the recipient is completely powerless (Handgraaf et al., 2008). And trying to impose constraints on power can backfire if it is targeted at those with less power (Foulk et al., 2020): If they perceive their position to be unstable or threatened, they trust others even less (Fast et al., 2012) and try to maintain power at others expense (Williams, 2014; Maner and Mead, 2010). Even innocuous features such as perceived legitimacy can amplify selfish behavior (de Cremer and van Dijk, 2005). So while transparency highlights moral costs and social responsibility for the very powerful, those with less power may be to preoccupied with the potential repercussions for their own status (Fast et al., 2012) and stability for such considerations.

Our paper is not without limitations. One concern the reader might have is the external validity of our setting. Specifically, we study a situation where one single individual is **randomly** elevated to a position of power for the **duration of the experiment**, without proper **accountability**. Obviously, this setting is unrealistic and does not reflect typical situations in real life. In real-life settings, there is a selection mechanism in place to elevate a person into a position of power. These mechanisms might include voting, merit, and corruption. However, all these situations induce one methodological issue: self-selection. Our goal was to study how a random person, who did not thrive for power, uses their power. Strikingly, we find that even a random person uses their power hypocritically. This makes our setting, even though not fully reflecting real life, very informative and internally valid.

Further, we abstract from proper accountability mechanisms. We did so on purpose to capture all those cases where accountability is difficult to achieve (police officers, politicians, etc.), where behavior is very difficult to observe (credence goods), and where accountability is, in fact, missing (authoritarian regimes and dictatorships). Thus, our setting speaks to all those situations where there are no proper checks and balances in place.

[32] It is hypocritical as it enforces double standards without normative justification. Undercontribution is widely considered to violate moral and social norms (Cubitt et al., 2011; Kimbrough and Vostroknutov, 2016), even for those with punishment power in our design Hoeft et al. (2019).

Finally, we induce the position of power for the duration of the experiment. This might be considered unrealistic as, in real life, people in power are, at some point, relieved from their power, be it by aging out of office, being voted out, or being fired. However, we consider this not to be at odds with our setting. Our setting merely induces the punisher to be in their position for a sufficient time. Prior experiments did not allow for a sufficient development of power structures by constantly reallocating power. We, on the other hand, wanted to study how power is affecting behavior if power is given for a sufficiently long time (and 60 minutes might not be considered a very long time). Therefore, we believe that our setting does speak to a wide range of settings, where power is held for a considerable time and proper accountability is missing or infeasible.

Another concern the reader might have is that our transparency treatments change more than one thing at a time. While our design explicitly changes only one thing, namely whether non-punishers can observe the contribution of the punisher, it induces a set of possible mechanisms through which transparency might work. It might be that punishers anticipate that their behavior will be considered to violate norms and therefore reduce their abusive behavior. It might, however, also be that punishers use this transparency to lead by example. In Appendix A.5, we study the latter and find no evidence of a leadership effect but rather a strong effect of punishment. However, there still remain more than a handful of possible mechanisms through which transparency might function (e.g., social norms, guilt, shame, etc., see also section 2 for more detail). In this study, we do not aim to distinguish between the different proposed psychological mechanisms and rather want to study the transparency institution wholesale (in line with most of the literature). Future research should focus on disentangling these possible psychological mechanisms to understand better how exactly transparency works.

Finally, our definition of abuse of power might be questioned. We argue that contributing below the contribution of other group members, which are still punished, can be interpreted as an abuse of power. Still, other interpretations also come to mind. It could be argued that punishers just behave rationally by freeriding and punishing non-punishers. It could also be argued that punishers just optimize labor division. Specifically, punishers have to monitor others' contributions and decide how to punish them and, therefore, are allowed to contribute less. Thus, punishers undercontributing the norm is a consequence of fair labor division. Further, it could be argued that punishers just play by the rules of the experiment and enjoy their randomly assigned position of power.[33] While all the above interpretations might represent the punishers' reasoning and justifications, recent research on social norms has shown that undercontribution is widely considered to violate moral and social norms (Cubitt et al., 2011; Kimbrough and Vostroknutov, 2016). Further, Hoeft et al. (2019) show that punishers, in exactly our setting, contributing below their peers is considered socially very inappropriate by punishers, non-punishers, and also by independent third parties. Thus, to accommodate the findings from social norms research, we interpreted punishers' behavior as an abuse of power.

Despite of these limitations, our findings 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 hypocritical exploitation and transparency and investigate under what circumstances resistance is organized.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jebo.2024.02.003.

References

Abbink, K., Gangadharan, L., Handfield, T., Thrasher, J., 2017. Peer punishment promotes enforcement of bad social norms. Nat. Commun. 8 (1).
Abbink, K., Irlenbusch, B., Renner, E., 2002. An experimental Bribery game. J. Law Econ. Organ. 18 (2), 428–454.
Albrecht, F., Kube, S., Traxler, C., 2018. Cooperation and norm enforcement - the individual-level perspective. J. Public Econ. 165, 1–16.
Andersson, P.A., Erlandsson, A., Västfjäll, D., Tinghög, G., 2020. Prosocial and moral behavior under decision reveal in a public environment. J. Behav. Exp. Econ. 87, 101561.
Andreoni, J., Bernheim, D.B., 2009. Social image and the 50 - 50 norm: a theoretical and experimental analysis of audience effects. Econometrica 77 (5), 1607–1636.
Andreoni, J., Croson, R., 2008. Partners versus strangers: random rematching in public goods experiments. In: Plott, C.R., Smith, V.L. (Eds.), 1 edition. Handbook of Experimental Economics Results, vol. 1, Part 6, chapter 82. Elsevier, pp. 776–783.
Andreoni, J., Gee, L., 2012. Gun for hire: delegated enforcement and peer punishment in public goods provision. J. Public Econ. 96 (11), 1036–1046.
Azfar, O., Nelson, W.R., 2007. Transparency, wages, and the separation of powers: an experimental analysis of corruption. Public Choice 130 (3), 471–493.
Azrieli, Y., Chambers, C.P., Healy, P.J., 2018. Incentives in experiments: a theoretical analysis. J. Polit. Econ. 126 (4), 1472–1503.

[33] Further, it could be argued that punishers improve the situation of non-punishers relative to a situation without any punishment. In Appendix A.9, we discuss this issue in detail and argue to show that the situation of non-punishers does not improve.

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. J. Pers. Soc. Psychol. 105 (10), 1013–1037.
Baldassarri, D., Grossman, G., 2011. Centralized sanctioning and legitimate authority promote cooperation in humans. Proc. Natl. Acad. Sci. 108, 11023–11027.
Bauhr, M., Grimes, M., 2014. Indignation or resignation: the implications of transparency for societal accountability. Governance 27 (2), 291–320.
Bell, A., Fairbrother, M., Jones, K., 2018. Fixed and random effects models: making an informed choice. Qual. Quant. 53 (2), 1051–1074.
Bendahan, S., Zehnder, C., Pralong, F.P., Antonakis, J., 2015. Leader corruption depends on power and testosterone. Leadersh. Q. 26 (2), 101–122.
Bock, O., Baetge, I., Nicklisch, A., 2014. hroot: Hamburg registration and organization online tool. Eur. Econ. Rev. 71 (C), 117–120.
Bodnar, T.J., Salathhè, M., 2013. The social maintenance of cooperation through hypocrisy.
Bolle, F., Vogel, C., 2011. Power comes with responsibility - or does it? Public Choice 148 (3), 459–470.
Bradley, A., Lawrence, C., Ferguson, E., 2018. Does observability affect prosociality? Proc. R. Soc. Lond. B, Biol. Sci. 285 (1875), 20180116.
Buffat, J., Senn, J., 2018. Corruption, norm enforcement and cooperation. Working Paper Series. Department of Economics, University of Zurich.
Burton-Chellew, M.N., Guérin, C., 2021. Decoupling cooperation and punishment in humans shows that punishment is not an altruistic trait. Proc. R. Soc. Lond. B, Biol. Sci. 288 (1962), 20211611.
Carpenter, J., Matthews, P.H., 2009. What norms trigger punishment? Exp. Econ. 12 (3), 272–288.
Carpenter, J.P., 2007. The demand for punishment. J. Econ. Behav. Organ. 62 (4), 522–542.
Carpenter, J.P., Matthews, P.H., 2012. Norm enforcement: anger, indignation, or reciprocity? J. Eur. Econ. Assoc. 10 (3), 555–572.
Castillo, J.G., Hamman, J., 2020. Political accountability and democratic institutions: an experimental assessment. J. Exp. Political Sci., 1–17.
Castillo, J.G., Xu, Z.P., Zhang, P., Zhu, X., 2020. The effects of centralized power and institutional legitimacy on collective action. Soc. Choice Welf. 56 (2), 385–419.
Charness, G., Gneezy, U., Halladay, B., 2016. Experimental methods: pay one or pay all. J. Econ. Behav. Organ. 131, 141–150.
Chen, C., Ganapati, S., 2023. Do transparency mechanisms reduce government corruption? A meta-analysis. Int. Rev. Adm. Sci. 89 (1), 257–272.
Cinyabuguma, M., Page, T., Putterman, L., 2006. Can second-order punishment deter perverse punishment? Exp. Econ. 9 (3), 265–279.
Cordis, A.S., Warren, P.L., 2014. Sunshine as disinfectant: the effect of state freedom of information act laws on public corruption. J. Public Econ. 115, 18–36.
Cubitt, R.P., Drouvelis, M., Gächter, S., Kabalin, R., 2011. Moral judgments in social dilemmas: how bad is free riding? J. Public Econ. 95 (3–4), 253–264.
Dasgupta, U., Radoniqi, F., 2021. Republic of Beliefs: An Experimental Investigation.
de Cremer, D., van Dijk, E., 2005. When and why leaders put themselves first: leader behaviour in resource allocations as a function of feeling entitled. Eur. J. Soc. Psychol. 35 (4), 553–563.
Dubois, D., Rucker, D.D., Galinsky, A.D., 2015. Social class, power, and selfishness: when and why upper and lower class individuals behave unethically. J. Pers. Soc. Psychol. 108 (3), 436–449.
Ertan, A., Page, T., Putterman, L., 2009. Who to punish? Individual decisions and majority rule in mitigating the free rider problem. Eur. Econ. Rev. 53 (5), 495–511.
Faillo, M., Grieco, D., Zarri, L., 2013. Legitimate punishment, feedback, and the enforcement of cooperation. Games Econ. Behav. 77 (1), 271–283.
Falco, S.D., Magdalou, B., Masclet, D., Villeval, M.C., Willinger, M., 2020. Can shorter transfer chains and transparency reduce embezzlement? Rev. Behav. Econ. 7 (2), 103–143.
Falk, A., Fehr, E., Fischbacher, U., 2005. Driving forces behind informal sanctions. Econometrica 73 (6), 2017–2030.
Fast, N.J., Halevy, N., Galinsky, A.D., 2012. The destructive nature of power without status. J. Exp. Soc. Psychol. 48 (1), 391–394.
Fehr, E., Fischbacher, U., 2004. Third-party punishment and social norms. Evol. Hum. Behav. 25 (2), 63–87.
Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415 (6868), 137–140.
Filiz-Ozbay, E., Ozbay, E.Y., 2014. Effect of an audience in public goods provision. Exp. Econ. 17 (2), 200–214.
Fischbacher, U., 2007. z-tree: Zurich toolbox for ready-made economic experiments. Exp. Econ. 10 (2), 171–178.
Fischbacher, U., Gachter, S., 2010. Social preferences, beliefs, and the dynamics of free riding in public goods experiments. Am. Econ. Rev. 100 (1), 541–556.
Fischbacher, U., Gächter, S., Fehr, E., 2001. Are people conditionally cooperative? Evidence from a public goods experiment. Econ. Lett. 71 (3), 397–404.
Fitzmaurice, G.M., Laird, N.M., 2014. Applied Longitudinal Analysis. Lightning Source Log In.
Foulk, T.A., Chighizola, N., Chen, G., 2020. Power corrupts (or does it?): an examination of the boundary conditions of the antisocial effects of experienced power. Soc. Pers. Psychol. Compass 14 (4), e12524.
Fox, J., 2007. The uncertain relationship between transparency and accountability. Dev. Pract. 17 (4–5), 663–671.
Fried, B.J., Lagunes, P., Venkataramani, A., 2010. Corruption and inequality at the crossroad: a multimethod study of bribery and discrimination in Latin America. Lat. Am. Res. Rev. 45 (1), 76–97.
Galinsky, A.D., Gruenfeld, D.H., Magee, J.C., 2003. From power to action. J. Pers. Soc. Psychol. 85 (3), 453–466.
Galinsky, A.D., Rucker, D.D., Magee, J.C., 2015. Power: past findings, present considerations, and future directions. In: APA Handbook of Personality and Social Psychology, vol. 3: Interpersonal Relations. American Psychological Association, Washington, DC, US, pp. 421–460.
Garcia-Gallego, A., Georgantzis, N., Jaber-Lopez, T., Michailidou, G., 2020. Audience effects and other-regarding preferences against corruption: experimental evidence. J. Econ. Behav. Organ. 180, 159–173.
Gelman, A., Hill, J., 2013. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.
Giurge, L.M., van Dijke, M., Zheng, M.X., De Cremer, D., 2021. Does power corrupt the mind? The influence of power on moral reasoning and self-interested behavior. Leadersh. Q. 32 (4), 101288.
Grimmelikhuijsen, S., Porumbescu, G., Hong, B., Im, T., 2013. The effect of transparency on trust in government: a cross-national comparative experiment. Public Adm. Rev. 73 (4), 575–586.
Handgraaf, M.J.J., Van Dijk, E., Vermunt, R.C., Wilke, H.A.M., De Dreu, C.K.W., 2008. Less power or powerless? Egocentric empathy gaps and the irony of having little versus no power in social decision making. J. Pers. Soc. Psychol. 95 (5), 1136–1149.
Hanna, P.R., Bishop, S., Nadel, S., Scheffler, G., Durlacher, K., 2010. The Effectiveness of Anti-Corruption Policy: What Has Worked, What Hasn't, and What We Don't Know. Short Works.
Hauge, K.E., 2015. Moral opinions are conditional on the behavior of others. Rev. Soc. Econ. 73 (2), 154–175.
Heisig, J.P., Schaeffer, M., Giesecke, J., 2017. The costs of simplicity: why multilevel models may benefit from accounting for cross-cluster differences in the effects of controls. Am. Sociol. Rev. 82 (4), 796–827.
Herrmann, B., Thöni, C., 2009. Measuring conditional cooperation: a replication study in Russia. Exp. Econ. 12 (1), 87–92.
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. Proc. Natl. Acad. Sci. 111 (2). 201315273–756.
Hoeft, L., Kurschilgen, M., Mill, W., Vannuccini, S., 2022. Norms as Obligations. Munich Papers in Political Economy 22. Munich School of Politics and Public Policy and the School of Management at the Technical University of Munich.
Hoeft, L., Mill, W., 2017. Selfish punishers. Econ. Lett. 157, 41–44.
Hoeft, L., Mill, W., Vostroknutov, A., 2019. Normative perception of power abuse.
Hood, C., Heald, D. (Eds.), 2006. Transparency: The Key to Better Governance, illustrated edition. British Academy, Oxford; New York.
Houser, D., Xiao, E., 2010. Inequality-seeking punishment. Econ. Lett. 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. Econ. Lett. 102 (3), 192–194.

Joshi, A., 2013. Do they work? Assessing the impact of transparency and accountability initiatives in service delivery. Dev. Policy Rev. 31, s29–s48.
Khadjavi, M., Lange, A., Nicklisch, A., 2017. How transparency may corrupt – experimental evidence from asymmetric public goods games. J. Econ. Behav. Organ. 142, 468–481.
Kimbrough, E.O., Vostroknutov, A., 2016. Norms make preferences social. J. Eur. Econ. Assoc. 14 (3), 608–638.
Kipnis, D., 1972. Does power corrupt? J. Pers. Soc. Psychol. 24 (1), 33–41.
Kirchkamp, O., Mill, W., 2021. Spite vs. risk: explaining overbidding in the second-price all-pay auction. Games Econ. Behav. 130, 616–635.
Kirchkamp, O., Mill, W., 2020. Conditional cooperation and the effect of punishment. J. Econ. Behav. Organ. 174, 150–172.
Klitgaard, R., 1988. Controlling corruption. In: Controlling Corruption. University of California Press.
Klitzman, R., 2007. When Doctors Become Patients. Oxford University Press Inc.
Kochel, T.R., Skogan, W.G., 2021. Accountability and transparency as levers to promote public trust and police legitimacy: findings from a natural experiment. Policing 44 (6), 1046–1059.
Kocher, M., Cherry, T., Kroll, S., Netzer, R.J., Sutter, M., 2008. Conditional cooperation on three continents. Econ. Lett. 101 (3), 175–178.
Kolstad, I., Wiig, A., 2009. Is transparency the key to reducing corruption in resource-rich countries? World Dev. 37 (3), 521–532.
Kosack, S., Fung, A., 2014. Does transparency improve governance? Annu. Rev. Pol. Sci. 17 (1), 65–87.
Kosfeld, M., Rustagi, D., 2015. Leader punishment and cooperation in groups: experimental field evidence from commons management in Ethiopia. Am. Econ. Rev. 105 (2), 747–783.
Kurzban, R., DeScioli, P., 2008. Reciprocity in groups: information-seeking in a public goods game. Eur. J. Soc. Psychol. 38 (1), 139–158.
Kurzban, R., DeScioli, P., O'Brien, E., 2007. Audience effects on moralistic punishment. Evol. Hum. Behav. 28 (2), 75–84.
Kuwabara, K., 2015. How does status affect power use? New perspectives from social psychology. In: Advances in Group Processes. Emerald Group Publishing Limited, pp. 99–121.
Kuwabara, K., Yu, S., 2017. Costly punishment increases prosocial punishment by designated punishers: power and legitimacy in public goods games. Soc. Psychol. Q. 80 (2), 174–193.
Lacetera, N., Macis, M., 2010. Social image concerns and prosocial behavior: field evidence from a nonlinear incentive scheme. J. Econ. Behav. Organ. 76 (2), 225–237.
Lammers, J., Galinsky, A.D., Dubois, D., Rucker, D.D., 2015. Power and morality. Curr. Opin. Psychol. 6, 15–19.
Lammers, J., Galinsky, A.D., Gordijn, E.H., Otten, S., 2012. Power increases social distance. Soc. Psychol. Pers. Sci. 3 (3), 282–290.
Lammers, J., Stapel, D.A., Galinsky, A.D., 2010. Power increases hypocrisy: moralizing in reasoning, immorality in behavior. Psychol. Sci. 21 (5), 737–744.
Leibbrandt, A., López-Pérez, R., 2012. An exploration of third and second party punishment in ten simple games. J. Econ. Behav. Organ. 84 (3), 753–766.
Malesky, E., Schuler, P., Tran, A., 2012. The adverse effects of sunshine: a field experiment on legislative transparency in an authoritarian assembly. Am. Polit. Sci. Rev. 106 (4), 762–786.
Mallucci, P., Wu, D.Y., Cui, T.H., 2019. Social motives in bilateral bargaining games: how power changes perceptions of fairness. J. Econ. Behav. Organ. 166, 138–152.
Maner, J.K., Mead, N.L., 2010. The essential tension between leadership and power: when leaders sacrifice group goals for the sake of self-interest. J. Pers. Soc. Psychol. 99 (3), 482–497.
Marcus, D.K., Zeigler-Hill, V., Mercer, S.H., Norris, A.L., 2014. The psychology of spite and the measurement of spitefulness. Psychol. Assess. 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. Rev. Econ. Stud. 81 (1), 301–324.
Markussen, T., Sharma, S., Singhal, S., Tarp, F., 2021. Inequality, institutions and cooperation. Eur. Econ. Rev. 138, 103842.
Mieth, L., Buchner, A., Bell, R., 2021. Moral labels increase cooperation and costly punishment in a prisoner's dilemma game with punishment option. Sci. Rep. 11 (1), 10221.
Milinski, M., Semmann, D., Krambeck, H.-J., 2002. Reputation helps solve the 'tragedy of the commons'. Nature 415 (6870), 424–426.
Mill, W., Morgan, J., 2021. The cost of a divided America: an experimental study into destructive behavior. Exp. Econ.
Mill, W., Morgan, J., 2022. Competition between friends and foes. Eur. Econ. Rev. 147, 104171.
Mill, W., Theelen, M.M., 2019. Social value orientation and group size uncertainty in public good dilemmas. J. Behav. Exp. Econ. 81, 19–38.
Molina, E., Carella, L., Pacheco, A., Cruces, G., Gasparini, L., 2016. Community monitoring interventions to curb corruption and increase access and quality of service delivery in low- and middle-income countries: a systematic review. Campbell Syst. Rev. 12 (1), 1–204.
Mooijman, M., van Dijk, W.W., Ellemers, N., van Dijk, E., 2015. Why leaders punish: a power perspective. J. Pers. Soc. Psychol. 109 (1), 75–89.
Murphy, R.O., Ackerman, K.A., 2014. Social value orientation: theoretical and measurement issues in the study of social preferences. Personal. Soc. Psychol. Rev. 18 (1), 13–41.
Murphy, R.O., Ackerman, K.A., Handgraaf, M.J.J., 2011. Measuring social value orientation. Judgm. Decis. Mak. 6 (8), 771–781.
Nikiforakis, N., Normann, H.-T., 2008. A comparative statics analysis of punishment in public-good experiments. Exp. Econ. 11 (4), 358–369.
Nikiforakis, N., Noussair, C.N., Wilkening, T., 2012. Normative conflict and feuds: the limits of self-enforcement. J. Public Econ. 96 (9–10), 797–807.
O'Gorman, R., Henrich, J., Van Vugt, M., 2009. Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. Proc. R. Soc. Lond. B, Biol. Sci. 276 (1655), 323–329.
Olken, B.A., Pande, R., 2012. Corruption in developing countries. Annu. Rev. Econ. 4 (1), 479–509.
Ostrom, E., Walker, J., Gardner, R., 1992. Covenants with and without a sword: self-governance is possible. Am. Polit. Sci. Rev. 86 (2), 404–417.
Otten, K., Buskens, V., Przepiorka, W., Ellemers, N., 2020. Heterogeneous groups cooperate in public good problems despite normative disagreements about individual contribution levels. Sci. Rep. 10 (1).
Parra, D., Munoz-Herrera, M., Palacio, L.A., 2021. The limits of transparency in reducing corruption. J. Behav. Exp. Econ. 95, 101762.
Peisakhin, L., 2012. Transparency and corruption: evidence from India. J. Law Econ. 55 (1), 129–149.
Peisakhin, L., Pinto, P., 2010. Is transparency an effective anti-corruption strategy? Evidence from a field experiment in India. Regul. Gov. 4 (3), 261–280.
Piazza, J., Bering, J.M., 2008a. Concerns about reputation via gossip promote generous allocations in an economic game. Evol. Hum. Behav. 29 (3), 172–178.
Piazza, J., Bering, J.M., 2008b. The effects of perceived anonymity on altruistic punishment. Evol. Psychol. 6 (3), 147470490800600314.
Putterman, L., Tyran, J.-R., Kamei, K., 2011. Public goods and voting on formal sanction schemes. J. Public Econ. 95 (9–10), 1213–1222.
Rus, D., van Knippenberg, D., Wisse, B., 2012. Leader power and self-serving behavior: the moderating role of accountability. Leadersh. Q. 23 (1), 13–26.
Rustagi, D., Engel, S., Kosfeld, M., 2010. Conditional cooperation and costly monitoring explain success in forest commons management. Science 330 (6006), 961–965.
Rustichini, A., Villeval, M.C., 2014. Moral hypocrisy, power and social preferences. J. Econ. Behav. Organ. 107, 10–24.
Salmon, T.C., Serra, D., 2017. Corruption, social judgment and culture: an experiment. J. Econ. Behav. Organ. 142, 64–78.
Sassenberg, K., Ellemers, N., Scheepers, D., 2012. The attraction of social power: the influence of construing power as opportunity versus responsibility. J. Exp. Soc. Psychol. 48 (2), 550–555.
Sassenberg, K., Ellemers, N., Scheepers, D., Scholl, A., 2014. "Power corrupts" revisited: the role of construal of power as opportunity or responsibility. In: Prooijen, J.-W.v., Lange, P.A.M.v. (Eds.), Power, Politics, and Paranoia: Why People Are Suspicious of Their Leaders. Cambridge University Press, Cambridge, pp. 73–88.
Schier, U.K., Ockenfels, A., Hofmann, W., 2016. Moral values and increasing stakes in a dictator game. J. Econ. Psychol. 56, 107–115.
Schilke, O., Reimann, M., Cook, K.S., 2015. Power decreases trust in social exchange. Proc. Natl. Acad. Sci. 112 (42), 12950–12955.
Senci, C.M., Hasrun, H., Moro, R., Freidin, E., 2019. The influence of prescriptive norms and negative externalities on bribery decisions in the lab. Ration. Soc. 31 (3), 287–312.

Snijders, T.A.B., Bosker, R., 2011. Multilevel Analysis. Sage Publications Ltd.
Sutter, M., Haigner, S., Kocher, M.G., 2010. Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations. Rev. Econ. Stud. 77 (4), 1540–1566.
Tost, L.P., Wade-Benzoni, K.A., Johnson, H.H., 2015. Noblesse oblige emerges (with time): power enhances intergenerational beneficence. Organ. Behav. Hum. Decis. Process. 128, 61–73.
Traulsen, A., Röhl, T., Milinski, M., 2012. An economic experiment reveals that humans prefer pool punishment to maintain the commons. Proc. R. Soc. Lond. B, Biol. Sci. 279 (1743), 3716–3721.
Vadlamannati, K.C., Cooray, A., 2017. Transparency pays? Evaluating the effects of the freedom of information laws on perceived government corruption. J. Dev. Stud. 53 (1), 116–137.
van Dijk, E., De Cremer, D., Handgraaf, M.J.J., 2004. Social value orientations and the strategic use of fairness in ultimatum bargaining. J. Exp. Soc. Psychol. 40 (6), 697–707.
van Kleef, G.A., Wanders, F., Stamkou, E., Homan, A.C., 2015. The social dynamics of breaking the rules: antecedents and consequences of norm-violating behavior. Curr. Opin. Psychol. 6, 25–31.
Vredenburgh, D., Brender, Y., 1998. The hierarchical abuse of power in work organizations. J. Bus. Ethics 17 (12), 1337–1347.
Weber, T.O., Weisel, O., Gächter, S., 2018. Dispositional free riders do not free ride on punishment. Nat. Commun. 9 (1).
Williams, M.J., 2014. Serving the self from the seat of power: goals and threats predict leaders' self-interested behavior. J. Manag. 40 (5), 1365–1395.
Wiltermuth, S.S., Flynn, F.J., 2013. Power, moral clarity, and punishment in the workplace. Acad. Manag. J. 56 (4), 1002–1023.
Wong, K.C., 1998. A reflection on police abuse of power in the People's Republic of China. Police Quarterly 1 (2), 87–112.
Wu, J., Balliet, D., Van Lange, P.A.M., 2016. Reputation, gossip, and human cooperation. Soc. Pers. Psychol. Compass 10 (6), 350–364.
Xiao, E., 2013. Profit-seeking punishment corrupts norm obedience. Games Econ. Behav. 77 (1), 321–344.
Xu, A.J., Loi, R., Lam, L.W., 2015. The bad boss takes it all: how abusive supervision and leader–member exchange interact to influence employee silence. Leadersh. Q. 26 (5), 763–774.
Zhang, B., Li, C., Silva, H., Bednarik, P., Sigmund, K., 2014. The evolution of sanctioning institutions: an experimental approach to the social contract. Exp. Econ. 17 (2), 285–303.
Zhou, Y., Jiao, P., Zhang, Q., 2017. Second-party and third-party punishment in a public goods experiment. Appl. Econ. Lett. 24 (1), 54–57.