



This postprint was originally published by Hogrefe as:
Dong, M., van Prooijen, J.-W., & van Lange, P. A. M. (2022).
**Strategic exploitation by higher-status people incurs harsher
third-party punishment.** *Social Psychology*, 53(4), 209–220.
<https://doi.org/10.1027/1864-9335/a000493>

Supplementary material to this article is available. For more information see
<https://hdl.handle.net/21.11116/0000-000B-69BF-9>

The following copyright notice is a publisher requirement:

This article version does not fully correspond to the article published in the journal *Social Psychology* under <https://doi.org/10.1027/1864-9335/a000493>. This is not the original version of the article and therefore cannot be used for citation.

Terms of use:

Available under the [CC BY-4.0 license](https://creativecommons.org/licenses/by/4.0/).



Provided by:

Max Planck Institute for Human Development
Library and Research Information
library@mpib-berlin.mpg.de

Strategic Exploitation by Higher-Status People Incurs Harsher Third-Party Punishment

Mengchen Dong^{1,2}, Jan-Willem van Prooijen¹, Paul A. M. van Lange¹

¹ Vrije Universiteit Amsterdam

² Max Planck Institute for Human Development

Corresponding author: Mengchen Dong (dong@mpib-berlin.mpg.de)

Author Notes

This research was supported by China Scholarship Council under grant (number: 201606040158). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors declare no conflict of interest.

STATUS, EXPLOITATION, AND PUNISHMENT

Abstract

It is widely documented that third parties punish norm violations, even at a substantial cost to themselves. However, little is known about how third-party punishment occurs in groups consisting of members that differ in status. Having a higher-status member promotes norm enforcement and group efficiency, but also poses threats to collective goods when they strategically exploit people's trust to maximize self-interest. Two pre-registered studies consistently revealed a punitive mechanism contingent on target status and strategic exploitation. Third-party observers generated harsher punishment when high- but not low-status targets transgressed after publicly endorsing cooperation (Study 1) or procedural fairness (Study 2). The findings elucidate third-party punishment as a feasible mechanism to counteract exploitation and maintain social norms in interactions with status asymmetry.

Keywords: social norm; norm violation; social status; deception; punishment

STATUS, EXPLOITATION, AND PUNISHMENT

Strategic Exploitation by Higher-Status People Incurs Harsher Third-Party Punishment

1 Introduction

Status asymmetry, characterized by imbalanced interpersonal esteem and respect, prevails across organizations, societies, and cultural groups (Henrich & Gil-White, 2001; Magee & Galinsky, 2008). High-status holders are usually expected to play an exemplary role in stimulating norm abidance of lower-status people. However, every now and then, they do exploit endowed trust by, for example, communicating social norms strategically to serve their self-interest. As a case in point, Dutch minister Ferdinand Grapperhaus was exposed of violating the COVID-19 social distancing rules at his own wedding celebration. Ironically, he was the exact person who oversaw justice issues related to coronavirus and had condemned violators of the 1.5-meter social distancing rules as “asociaal” (the Dutch equivalent of “antisocial”). In ambivalent times like the COVID-19 pandemic, status holders may not strictly comply with normative regulations, but can gain reputational benefits and stabilize their high-status positions (e.g., being seen as responsible and adequate leaders) through advocating appropriate norms. Then, how do people counteract such strategic exploitation of high-status members? And how are social norms eventually sustained in status-asymmetrical relationships?

Governments and organizations are often urged to increase civic engagement in leadership ethics regulations, whereas it is still under debate whether people are capable of detecting exploitative leaders and hold them accountable for their actions (Bøggild, 2020; DeScioli & Bokemper, 2019; Padilla et al., 2007). On the one hand, people are motivated to imitate status holders and deem their behaviors as adequate normative practices (including selfish or unethical ones; Bauman et al., 2016; Bunderson, 2003; Simpson et al., 2012). On the other hand, high-status holders' violations of communicated norms introduce more

STATUS, EXPLOITATION, AND PUNISHMENT

difficulty to maintain member commitment, mutual trust, and group identity (Abrams et al., 2013; Bauman et al., 2016; Simons et al., 2015).

The current research examines whether third-party punishment is a viable mechanism to enforce social norms in the presence of status-based strategic exploitation. Third-party punishment serves an important function to maintain social norms (Carlsmith et al., 2002; Fehr & Fischbacher, 2004; McAuliffe et al., 2015; for a meta-analytical review, see Balliet et al., 2011). In particular, compared to people who pursue self-interest without using strategic communications, strategic exploiters can be seen as worse (Jordan et al., 2017; Ohtsubo et al., 2010), especially when they have high (vs. low) status and are expected to do as they say (Dong et al., 2021). We therefore advance the general hypothesis that high- (but not low-) status holders elicit stronger third-party punishment when they are strategic exploiters rather than openly selfish actors. Below, we present our reasoning in greater detail.

1.1 Third-Party Punishment of Strategic Exploitation

Substantial research suggests that independent third parties are willing to sacrifice personal resources to punish norm violators (Balliet et al., 2011; Fehr & Fischbacher, 2004; Henrich et al., 2006; McAuliffe et al., 2015). The extent of third-party punishment also increases with the severity of transgressions (Boyd et al., 2003; Carlsmith et al., 2002; Fehr & Fischbacher, 2004; Van Prooijen, 2018). Selfishness and deception both violate social norms, and elicit third-party punishment to maintain cooperation and sincerity norms (e.g., Fehr & Fischbacher, 2004; Ohtsubo et al., 2010). We therefore reason that people should punish strategic exploiters extremely harshly since it reflects both selfishness and deception. We conceptualize strategic exploitation as being selfish while sending misleading messages. For instance, in a prisoner's dilemma, strategic exploiters can defect while sending a message designed to mislead their partners to cooperate, which yields the best material payoffs for themselves. Or, in a dictator game, strategic exploiters can make a selfish distribution while

STATUS, EXPLOITATION, AND PUNISHMENT

sending a message to mislead their partners to blame the outcome on an ostensibly objective procedure (e.g., anonymous dice rolling), which maintains their fair reputation and deters partner retribution (Lönqvist et al., 2015). In addition to a selfish behavior, prior misleading communications can incur harsher punishment given their deceptive nature (Boles et al., 2000; Brandts & Charness, 2003; Ohtsubo et al., 2010). People may deceive for self-oriented or other-oriented reasons, but only self-oriented deception receives harsh third-party punishment (Levine & Schweitzer, 2014, 2015). Strategic exploitation is often associated with self-oriented motives (Jordan et al., 2017; Dong et al., 2021), and therefore should incur third-party punishment, which is stronger than being openly selfish or exploitative.

We reason that people sanction strategic exploiters to enforce social norms and maintain social order. However, there might be other mechanisms that drive third-party punishment of strategic exploitation. For example, some recent studies provide a second explanation, suggesting that people punish transgressors to defend self-benefits in potential interactions with the transgressor (Delton et al., 2011; Krasnow et al., 2016). People make analogous inferences about how the target will treat themselves from how the target treats others. Therefore, third parties punish the targets who mistreat others to avoid potential mistreatment of themselves (Krasnow et al., 2016). A third explanation is also possible, which focuses on the reputational benefits of third-party punishers (Jordan et al., 2016; Nelissen, 2008). Third-party costly punishment can signal the punishers' trustworthiness and facilitate their benign interactions with those who know their punitive actions (Jordan et al., 2016).

The three explanations differ in terms of the social motivation that promotes punishment. Our first explanation implies a non-selfish mechanism, in that punishment derives from concerns for collective goods regardless of whether one can gain from such costly behavior. In contrast, the two alternative explanations suggest selfish motives

STATUS, EXPLOITATION, AND PUNISHMENT

underlying third-party punishment, through which people expect to either gain a good reputation in the eyes of beholders, or to prevent future loss in interactions with the punished transgressor. In the current research, in addition to examining the hypothesized behavioral punishment of strategic exploitation, we will also explore the underlying self- versus other-oriented mechanisms.

1.2 The Role of Status

While much is known about third-party punishment in interactions with equal status, it is not yet clear how it works in groups with status asymmetry. We conceptualize status as social prestige and esteem afforded by others (Henrich & Gil-White, 2001; Magee & Galinsky, 2008). People are more likely to trust high- as opposed to low-status members when both communicate social norms (Barkow, 2014; Bunderson, 2003; Henrich & Gil-White, 2001; Simpson et al., 2012). Having a high-status member thus contributes to group performance and norm enforcement (Barkow, 2014; Bunderson, 2003; Simpson et al., 2012). However, it also provides the high-status person with privileged opportunities to exploit others. Status holders can exploit social trust and communicate strategically and deceptively, to maximize self-interest or consolidate their reputation (Case et al., 2018; Hays & Blader, 2017).

Little is known about how groups counteract status holders' strategic exploitation. In the current research, we investigate third-party sanctions as a protective tool against strategic exploitation of status holders, and expect particularly high levels of costly punishment when high- but not low-status targets transgress while preaching a contradictory norm. Since status holders are often expected to have behavioral integrity and practice what they preach (Henrich & Gil-White, 2001), their violation of communicated norms can be perceived as more deceptive and severe (Dong et al., 2021). Moreover, high-status members' transgressions with (vs. without) strategic communications can be especially detrimental to

STATUS, EXPLOITATION, AND PUNISHMENT

partner and group outcomes. Partners and followers (but not third-party observers; Risen & Gilovich, 2007) are motivated to trust, and defer to, higher-status targets' communications (Barkow, 2014; Henrich & Gil-White, 2001; Simpson et al., 2012), and thus incur more loss when high-status members exploit such communications. Finally, as high-status members are also expected to represent a group and stimulate norm abidance of lower-status people, their deceptive communications jeopardize normative group functioning to a greater extent by introducing difficulty to restore mutual trust, group identity, and member commitment (Abrams et al., 2013; Bauman et al., 2016; Simons et al., 2015). Thus, status holders' strategic exploitations should be seen as more deceptive and more detrimental to the group, and thus should receive extremely harsh third-party punishment.

Based on the above line of reasoning, we expect an interaction effect of transgression type (selfish choices with versus without strategic communication) and transgressor status (high versus low). In particular, as compared to actors who openly pursue self-interest, strategic exploiters should incur harsher third-party punishment for their deceptive communications of non-selfish social norms (e.g., cooperation or fairness), particularly when they possess high rather than low social status. Nonetheless, as people often justify status holders' behaviors as adequate normative practices (Bauman et al., 2016; Bunderson, 2003; Simpson et al., 2012), we cannot fully exclude the possibility that people give high-status strategic exploiters more leniency than their low-status counterparts.

1.3 Overview of the Present Research

In the present research, we conducted two pre-registered experiments (see the pre-registration forms at <https://aspredicted.org/vr9hc.pdf> and <https://aspredicted.org/29d4x.pdf>, to examine how third-party observers sacrifice their own endowment to punish norm violations, depending on the status of transgressors and their use of strategical communications before acting on their self-interest. Previous research showed that less than

STATUS, EXPLOITATION, AND PUNISHMENT

20% of the participants would deceive (e.g., send a cooperative message and then defect) in real interactions where there were risks of being punished (Ohtsubo et al., 2010); therefore, we adopt the *strategy method*, in which participants make conditional decisions for each possible outcome.

Study 1 manipulated status as relative esteem of the transgressors as compared to their partners, while Study 2 manipulated status broadly as general social prestige. Moreover, strategic exploitation was conceptualized through both material (Study 1) and reputational (Study 2) benefits. Specifically, we operationalized strategic exploitation as defection after sending a cooperative norm message in Study 1, and as transgressing a self-proclaimed fair procedure in Study 2. In addition to our main hypothesis, we additionally explored the roles of selfish (i.e., to defend self-benefit in potential interactions with the transgressor; Study 1) and non-selfish (i.e., to enforce social norms; Studies 1 and 2) motives in third-party punishment of status-based strategic exploitation. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the studies. The data, analysis code, and experimental materials can be accessed at shorturl.at/jknU5.

2 Study 1

Study 1 examined costly punishment of high- (vs. low-) status targets' strategic exploitation in a prisoner's dilemma. Participants (who always had the role of a third-party observer) were informed that the relatively higher- (vs. lower-) status person was assigned to a Sender role and could send a message to their relatively lower- (higher-) status partner as a Receiver. We expected particularly harsh third-party punishment when the high- (but not low-) status Sender defects after sending a message to advocate cooperation to the Receiver.

2.1 Method

2.1.1 Participants and design

STATUS, EXPLOITATION, AND PUNISHMENT

We employed a 2 (Sender status: high versus low) by 2 (Message: sent versus not sent) by 2 (Sender choice: cooperation versus defection) by 2 (Receiver choice: cooperation versus defection) mixed design with only Sender status as a between-participants factor. A priori power analysis yielded a sample of $N = 198$ to detect a small behavior by status interaction effect ($\eta_p^2 = .01$, $\alpha = 0.05$, with 80% power). Therefore, we intended to recruit 200 participants as third-party observers from the online crowdsourcing platform Turkprime.com (Litman et al., 2017). Two hundred and two participants (116 males; $M_{\text{age}} = 38.7$ years, $SD = 11.0$) passed the comprehension questions inserted throughout the instructions, and all of them were included upon their completion of the survey. To increase realism of the design, participants were matched with Senders and Receivers from a previous study, among which 5% of the three-person groups received the actual bonus as decided by their matched decisions.

2.1.2 Procedure

All participants played the role of Observer, and were teamed up with two other persons, respectively as Sender and Receiver. Participants were randomly assigned to read about the Sender having higher or lower status than the Receiver, and decided whether they would punish the Sender and Receiver in each of the eight possible cases.

After knowing their own role as an Observer, participants were introduced to a one-shot sequential prisoner's dilemma between the Sender and the Receiver. Specifically, Cooperate-Defect yielded (\$0, \$4), Cooperate-Cooperate yielded (\$3, \$3), Defect-Cooperate yield (\$4, \$0), and Defect-Defect yielded (\$1, \$1). In addition to the basic rules, the Sender had a unique chance to send a predetermined message to the Receiver, prior to their choice between cooperation and defection. The message communicated a cooperation norm, saying "I think people should definitely cooperate in this game". The Receiver knew whether or not the Sender sent the message and then decided whether to cooperate or defect. Importantly,

STATUS, EXPLOITATION, AND PUNISHMENT

Observers were made clear that “the message may NOT necessarily bind the Sender’s choice. That is, the Sender is free to send the message and then choose to defect”, which was also made known to the Receiver.

Participants then completed an ostensible social status survey, and were assigned to either a high-status Sender ($n = 99$) or a low-status Sender ($n = 103$) group based on bogus feedback. In the high-status (versus low-status, in the brackets) Sender condition, participants read that “As compared to Receivers, Senders are those who are MORE (LESS) respected and held in HIGHER (LOWER) esteem. Receivers (Senders) usually look up the Senders (Receivers) and admire them. Thus, Senders can send a message and strongly influence Receivers’ choice of cooperation/defection (Receivers can choose whether to cooperate/defect after knowing Senders’ preference through the message). This gives Senders relatively HIGHER (LOWER) STATUS than Receivers.” Participants then answered three questions as a manipulation check of Sender status ($\alpha = .73$; e.g., “To what extent do you think Senders have lower or higher status than Receivers?” rated on a 7-point scale ranging from 1 = *Definitely lower status* to 7 = *Definitely higher status*) and indicated their moral expectation of the Sender (“To what extent do you expect Senders to be morally worse or better persons than Receivers?” rated on a 7-point scale ranging from 1 = *Morally worse* to 7 = *Morally better*).

Participants were eventually introduced to their own role with a \$3 endowment and could use the endowment to deduct the earnings of both the Sender and the Receiver at a 1:3 ratio (i.e., if the Observer pays \$0.1, the Sender or Receiver loses \$0.3). Observers responded to all the possible Sender-Receiver choice combinations in a random sequence. An example is shown below:

- *Sender has HIGHER STATUS than Receiver.*

STATUS, EXPLOITATION, AND PUNISHMENT

- *Sender sends the message saying “I think people should definitely cooperate in this game” and then chooses to DEFECT*
- *Receiver receives the message and then chooses to COOPERATE*

After reading about each of the eight situations, participants rated perceptions of self-/other-oriented motives (one item; i.e., “How selfish or generous do you think the Sender’s/Receiver’s reasons are in doing so?” rated from 1 = *Completely selfish* to 7 = *Completely generous*) and the amount of money they intended to pay to punish (on a 15-point scale, ranging from \$0 to \$1.5), targeting respectively at the Sender and the Receiver. In addition, we explored how Observers presumed their interaction with the depicted Sender in a dictator game. Participants were asked to imagine a different game played with the Sender, and indicate how much money of a \$10 endowment the Sender would transfer to them in an anonymous setting (on a 100-point scale, ranging from \$0 to \$10).

2.2 Results

2.2.1 Manipulation check

As intended, participants evaluated the Sender as higher on status than the same-group Receiver, $t(186) = 10.82, p < .001; d = 0.895$, in the high-status ($M = 5.30, SD = 1.02$) than low-status ($M = 3.44, SD = 1.40$) condition. Additionally, people expected higher-status ($M = 4.49, SD = 1.03$; vs. lower-status, $M = 4.12, SD = 0.88$) Senders to be morally better persons, $t(192) = 2.80, p = .006; d = 0.928$.

2.2.2 Costly punishment

The descriptive information of costly punishment in each condition can be found in Table 1. In particular, we examined the situations where the Sender defected with or without sending a message to communicate cooperation norms. Therefore, as pre-registered, we conducted a repeated-measure ANOVA, to examine how Sender status, message (sent versus not), and Receiver’s behavior (cooperation versus defection) influenced third-party

STATUS, EXPLOITATION, AND PUNISHMENT

punishment of Sender's defection. As predicted, we found a significant two-way interaction between Sender status and message (see Figure 1), $F(1, 200) = 11.21, p = .001, \eta_p^2 = .053$, such that people punished the defected Sender more harshly when sending (vs. not sending) a deceptive cooperation message, but only when the Sender possessed higher ($M_{\text{deception}} = 0.50, SD = 0.58$; versus $M_{\text{non-deception}} = 0.36, SD = 0.55$), $F(1, 200) = 20.42, p < .001, \eta_p^2 = .093$, instead of lower status than their partner (i.e., the Receiver; $M_{\text{deception}} = 0.32, SD = 0.57$; versus $M_{\text{non-deception}} = 0.32, SD = 0.54$), $F(1, 200) = 0.03, p = .86, \eta_p^2 < .001$. Importantly, the non-significant three-way interaction, $F(1, 200) = 0.53, p = .45, \eta_p^2 = .003$, suggested that people's harsher punishment of higher-status strategic exploiters was not influenced by the behavioral consequence (i.e., cooperation/defection of the Receiver).

Table 1.

The means and standard deviations (in brackets) of observer costly punishment (unit: US dollar) on the Sender and Receiver respectively.

		Sender			
Conditions		Message & Defect	Message & Cooperate	No Message & Defect	No Message & Cooperate
Punishment of Higher-Status Sender ($n = 99$)					
	Cooperate	0.51(0.49)	0.24(0.39)	0.36(0.45)	0.23(0.38)
	Defect	0.48(0.49)	0.25(0.41)	0.35(0.44)	0.25(0.40)
Punishment of Lower-Status Sender ($n = 103$)					
	Cooperate	0.36(0.42)	0.18(0.34)	0.38(0.40)	0.16(0.32)
Receiver	Defect	0.28(0.32)	0.16(0.32)	0.27(0.37)	0.17(0.31)
Punishment of Lower-Status Receiver ($n = 99$)					

STATUS, EXPLOITATION, AND PUNISHMENT

Cooperate	0.26(0.42)	0.23(0.40)	0.30(0.39)	0.23(0.41)
Defect	0.25(0.43)	0.46(0.47)	0.30(0.42)	0.40(0.43)
Punishment of Higher-Status Receiver ($n = 103$)				
Cooperate	0.15(0.42)	0.19(0.35)	0.17(0.32)	0.15(0.30)
Defect	0.26(0.34)	0.40(0.44)	0.26(0.38)	0.34 (0.40)

Besides, we found main effects of message, $F(1, 200) = 9.64, p = .002, \eta_p^2 = .046$, and Receiver choice, $F(1, 200) = 13.11, p < .001, \eta_p^2 = .061$, suggesting that Observers punished more when the defected Sender sent the deceptive cooperation message ($M = 0.41, SD = 0.41$) versus not ($M = 0.34, SD = 0.38$), and when the Receiver cooperated ($M = 0.40, SD = 0.40$) rather than defected ($M = 0.34, SD = 0.37$). A status by Receiver choice interaction effect also emerged, $F(1, 200) = 5.14, p = .03, \eta_p^2 = .025$, such that Observers paid more to punish the defected Sender when the Receiver cooperated instead of defected, while only when the Sender possessed lower ($M_{\text{cooperate}} = 0.37, SD = 0.57$; versus $M_{\text{defect}} = 0.27, SD = 0.51$), $F(1, 200) = 17.67, p < .001, \eta_p^2 = .081$, rather than higher status ($M_{\text{cooperate}} = 0.44, SD = 0.57$; versus $M_{\text{defect}} = 0.42, SD = 0.53$), $F(1, 200) = 0.90, p = .34, \eta_p^2 = .004$.

2.2.3 Potential Mechanisms

We explored two possible mechanisms with repeated-measure ANOVAs: (1) Perception of the Sender's self-/other-oriented motives; (2) Imagined potential interactions with the Sender. Consistent with the results on costly punishment, we found a Sender status by message two-way interaction effect on motive perception, $F(1, 200) = 4.85, p = .03, \eta_p^2 = .024$. As shown in Figure 1, Observers perceived stronger selfish motives only when the deceptive Sender possessed higher status ($M_{\text{deception}} = 2.71, SD = 2.37$; versus $M_{\text{non-deception}} = 2.89, SD = 2.32$), $F(1, 200) = 3.96, p = 0.05, \eta_p^2 = .019$, but not when they had relatively

STATUS, EXPLOITATION, AND PUNISHMENT

lower status ($M_{\text{deception}} = 2.63$, $SD = 2.33$; versus $M_{\text{non-deception}} = 2.53$, $SD = 2.27$), $F(1, 200) = 1.24$, $p = .27$, $\eta_p^2 = .006$. Despite the consistent pattern of costly punishment and motive perception as a function of Sender status by message interaction effect, a within-participants mediation analysis with 5,000 simulations (Tingley et al., 2014) did not support a significant mediation of motive perception (indirect effect = -0.003 , 95% CI $[-0.008, 0.00]$, $p = .27$) in the effect of status by message interaction on costly punishment (total effect = 0.037 , 95% CI $[0.07, 0.07]$, $p = .018$; direct effect = 0.040 , 95% CI $[0.011, 0.07]$, $p = .007$).

Regarding imagined interactions with the Sender, we found a significant interaction effect between Sender status and message (see also Figure 1), $F(1, 200) = 4.53$, $p = .04$, $\eta_p^2 = .022$. However, people did not presume the deceptive Senders as more selfish in interactions with themselves, regardless of the Sender's high ($M_{\text{deception}} = 1.83$, $SD = 3.28$; versus $M_{\text{non-deception}} = 1.68$, $SD = 3.20$), $F(1, 200) = 2.39$, $p = .12$, $\eta_p^2 = .012$, or low status ($M_{\text{deception}} = 1.83$, $SD = 3.23$; versus $M_{\text{non-deception}} = 1.96$, $SD = 3.13$), $F(1, 200) = 2.15$, $p = .15$, $\eta_p^2 = .011$.

STATUS, EXPLOITATION, AND PUNISHMENT

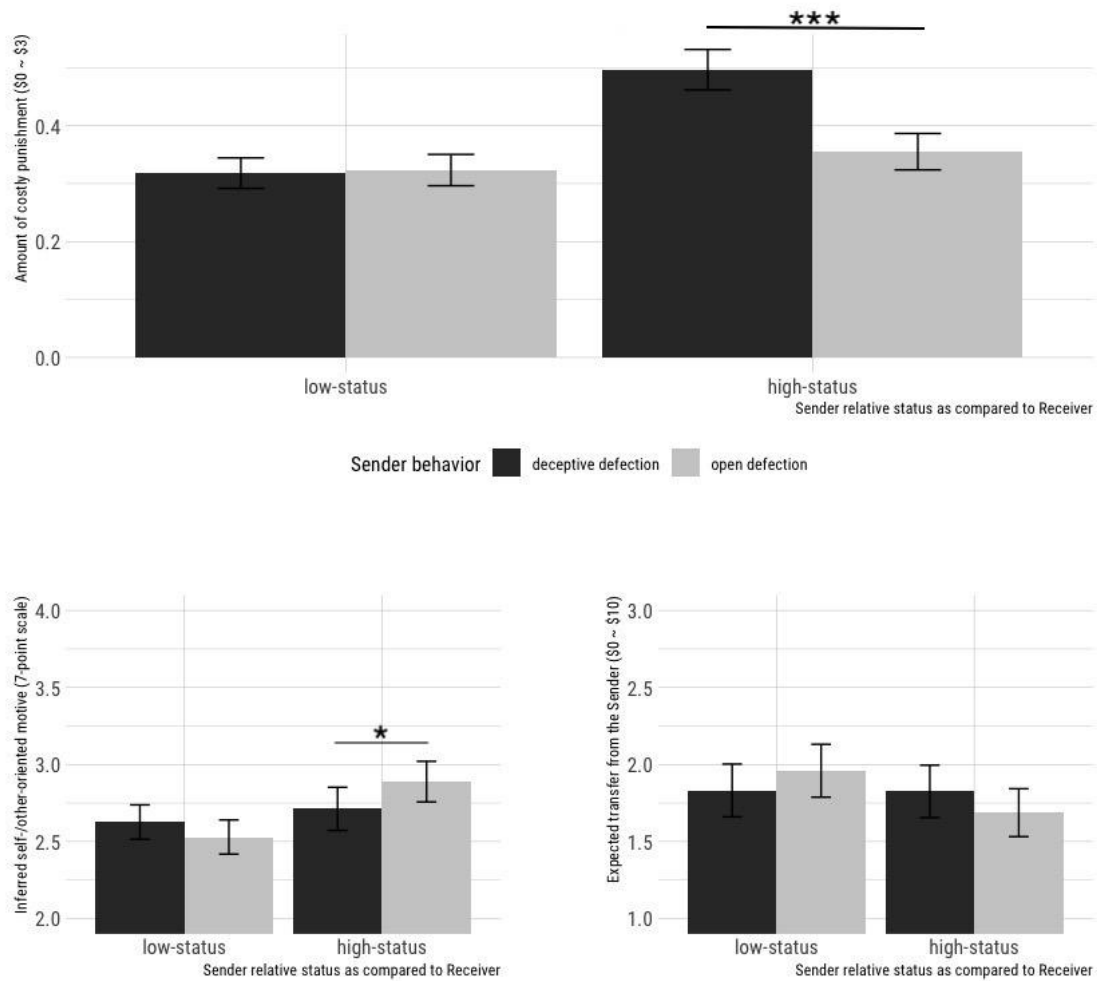


Figure 1. Observers' costly punishment (upper panel), inference of Sender motives (lower left panel; with a lower score representing more self-interested motives), and imagined money transferred from the Sender to the self (lower right panel), as a function of Sender status and behavior in Study 1. $*p < .05$. $***p < .001$.

2.3 Discussion

In a prisoner's dilemma, we found that third-parties' punitive decisions were influenced by defector status and their strategic exploitation. People punished high-status defectors more harshly only after the defectors sent a message to communicate a cooperative norm. Moreover, observers perceived stronger selfish motives from higher- (vs. lower-) status

STATUS, EXPLOITATION, AND PUNISHMENT

strategic exploiters, although these motives did not mediate the increased levels of punishment. Punishment also was unrelated to concerns about potential threats to self-interest in an imagined interaction.

3 Study 2

Study 2 aims to replicate the target status by strategic exploitation interaction effect in Study 1, with two main changes. First, we examined observer punishment in a dictator game where the strategic exploiters as Distributors advocated an objective procedure of dice guessing but cheated on it in private. In this case, strategic exploiters gained reputational benefits but not material benefits, as compared to openly selfish Distributors who assigned themselves with more money without deceptive communication. Second, we manipulated status more generally as societal prestige, instead of relative esteem between interaction partners. As in many real-life situations, high-status transgressors are originally esteemed social members, not only to the victims but also to uninvolved others. The esteem may come from various sources, such as perceptions of competence, courage, and past accomplishments which may partially be explained by some level of integrity. For these or related reasons, third-party observers may not punish high-status transgressors when they perceive the transgressors as having more authority than themselves (Yudkin et al., 2020). It is therefore important to replicate our findings and operationalize status broadly with relevance to third-party observers.

3.1 Method

3.1.1 Participants and design

Based on the power analysis in Study 1, we intended to recruit 200 participants as Observers from Turkprime.com. We employed a 2 (behavior: strategic versus open selfishness) by 2 (target status: high versus low) mixed design with target status as a

STATUS, EXPLOITATION, AND PUNISHMENT

between-participants factor. To eliminate experimental deception, we also recruited 200 Distributors online and matched them with another 200 Recipients from our previous study. As determined by their matched decisions, 5% of the Distributor-Recipient-Observer groups received the actual bonus. After passing some comprehension check questions, 193 participants as Observers (111 males; $M_{age} = 38.3$ years, $SD = 10.3$) completed our survey and were all included in further analysis.

3.1.2 Procedure

We mainly elaborate on the instructions for the Observer role below. Participants were informed to team up as three-person groups and play an interactive game. After being introduced to all the Roles (i.e., Distributor, Recipient, and Observer), participants read that their in-group Distributor had either high or low social status, and made punitive decisions for all the potential Distributor choices.

Participants were first introduced to the role of Distributor, who was asked to allocate \$10 with another Recipient. The Distributor had three options: (1) \$8 to the self and \$2 to the Recipient, (2) \$5 to the self and \$5 to the Recipient, or (3) guessing two six-sided dice online. If the Distributor chose to guess the dice, two dice—one in red and the other in green—appeared below. The Distributor then guessed two numbers *in their mind*, clicked on a button to throw the dice, and eventually indicated whether they guessed both the red and green dice correctly (i.e., Yes/No; see Figure 2 for screenshots of the interface). If the Distributor indicated to have guessed both the numbers correctly, he/she would get (1) \$8 to the self and \$2 to the Recipient. Instead, if the Distributor indicated to have guessed any of the two numbers incorrectly, he/she would get (2) \$5 to the self and \$5 to the Recipient. After the instructions above, all participants experienced the role of Distributor once as an exercise session. Second, we briefly introduced the role of Recipient, who made no active choice and only received the bonus (either \$2 or \$5) as determined by the Distributor's choice. Third,

STATUS, EXPLOITATION, AND PUNISHMENT

participants were introduced to the Observer role. The Observer had a \$5 endowment and could choose whether to use some of it to punish the Distributor at a 1:3 ratio (i.e., if the Observer pays \$0.1, the Distributor loses \$0.3). Given the minimal (i.e., 1/36) chance to actually guess both the dice numbers correctly, we conceptualized those who gave themselves \$8 (\$2 to the Recipients) through the ostensibly objective procedure of dice guessing as likely exploitative by strategically pursuing undeserved moral credits—as compared to those who directly gave themselves \$8 (\$2 to the Recipients) as openly selfish Distributors. To check participants' understanding of the scenario, we presented two questions (“Do Distributors have a high chance to win him-/herself \$8 if they guess the dice?” on a 7-point scale from 1 = *Definitely no* to 7 = *Definitely yes*; “If a Distributor claims to have guessed the dice correctly to win him-/herself \$8, how likely is he/she lying?” on a 7-point scale from 1 = *Very unlikely* to 7 = *Very likely*).

STATUS, EXPLOITATION, AND PUNISHMENT

How do you want to distribute \$10?

\$8 to the self and \$2 to a recipient

\$5 to the self and \$5 to a recipient

Guess two dice

How do you want to distribute \$10?



\$8 to the self and \$2 to a recipient

\$5 to the self and \$5 to a recipient

Guess two dice

GUESS:

What is the RED dice? What is the GREEN dice?

Throw the dice!!

Did you guess both the dice correctly?

YES! ---Congrats! You get \$8 for yourself and \$2 for a recipient

NO!--- Not bad! You get \$5 for yourself and \$5 for a recipient

Figure 2. Screenshot pictures of Distributor choices. Observers also made punitive decisions based on these pictures, with the upper picture representing an openly selfish choice and the lower picture representing an exploitative selfish choice.

Targeted participants were then informed about their own role as an Observer, and were randomly assigned to either a high-status ($n = 97$) or a low-status ($n = 96$) Distributor group, knowing that the Distributor scored either superior or moderate on a social status survey. Specifically, Observers in the high-status (versus low-status, in the bracket) group were informed that “Distributors are those who are HIGHLY respected and held in HIGH esteem as compared to most others around them (respected and held in esteem SIMILARLY

STATUS, EXPLOITATION, AND PUNISHMENT

as most people are). People look up to the Distributors and admire them TO A GREAT EXTENT (no more or less than average others), and this makes them very HIGH-STATUS persons (gives them a reasonably AVERAGE-STATUS) in their work and life.” After reading the Distributor status information, Observers evaluated their perception of Distributor status as a manipulation check (e.g., “In general, to what extent do you feel the Distributors are low-status or high-status persons?”; 1 = *Absolutely low-status* to 7 = *Absolutely high-status*; $\alpha = .75$ across the three items). Then, in a random sequence, participants were presented with screenshot pictures (as in Figure 2) indicating respective choices of open selfishness, strategic selfishness, and two other filler conditions. For each Distributor choice, participants indicated their (1) perceived self-/other-oriented motives (“How selfish or generous do you think the Distributor’s reasons are in doing so?” rated from 1 = *Completely selfish* to 7 = *Completely generous*; “Do you think that the Distributor does so because he/she cares more about doing what is the best for him-/herself versus what is the best for the Recipient?” rated from 1 = *Only about him-/herself* to 7 = *Only about the Recipient*; $r = .87$, $p < .001$) and (2) the amount of money they wanted to pay to punish the Distributor (on a 25-point scale, ranging from \$0 to \$2.5).

3.2 Results

3.2.1 Manipulation checks

We first checked how our manipulation of strategic selfishness worked. As intended, both as compared with the scale midpoint 4.0, participants perceived a low chance of winning \$8 to the Distributor (\$2 to the Recipient) through dice guessing ($M = 3.57$, $SD = 2.36$), $t(192) = -2.53$, $p = .01$, $d = 0.182$, and a high probability of deception herein ($M = 5.61$, $SD = 1.41$), $t(192) = 15.78$, $p < .001$, $d = 1.142$. Also as expected, high-status ($M = 5.57$, $SD = 1.05$; versus low-status, $M = 5.02$, $SD = 0.89$) Distributors were seen as significantly higher on social status, $t(191) = 3.91$, $p < .001$, $d = 0.57$.

STATUS, EXPLOITATION, AND PUNISHMENT

3.2.2 Costly punishment

As pre-registered, we conducted a 2 (status: high vs. low) by 2 (behavior: strategic vs. open selfishness) repeated-measure ANOVA with status as a between-participants and behavior as a within-participants factor¹. Neither the Distributors' status, $F(1, 191) = 0.46$, $p = .498$, $\eta_p^2 = .002$, nor their behavior, $F(1, 191) = 1.39$, $p = .240$, $\eta_p^2 = .007$, had a significant main effect on costly punishment. Importantly, as predicted, we found a significant status by behavior interaction effect, $F(1, 191) = 7.03$, $p = .01$, $\eta_p^2 = .035$ (see Figure 3), suggesting that people invested more money to punish deceptively than openly selfish Distributors, but only when the Distributors possessed high ($M_{\text{deception}} = 0.91$, $SD = 0.87$; versus $M_{\text{non-deception}} = 0.71$, $SD = 0.85$), $F(1, 191) = 7.37$, $p = .01$, $\eta_p^2 = .037$, rather than low social status ($M_{\text{deception}} = 0.74$, $SD = 0.86$; versus $M_{\text{non-deception}} = 0.78$, $SD = 0.89$), $F(1, 191) = 1.08$, $p = .30$, $\eta_p^2 = .006$.

3.2.3 Perception of self-/other-oriented motives

We then explored how inferred motives played a role in third-party punishment. With a lower score representing more self-oriented and less other-oriented motives, we found a status by behavior interaction effect (see Figure 3), $F(1, 191) = 99.16$, $p < .001$, $\eta_p^2 = .342$, such that participants perceived stronger self-interested motives from high-status Distributors' strategic exploitation ($M = 2.96$, $SD = 2.05$) than open selfishness ($M = 4.00$, $SD = 1.67$), $F(1, 191) = 69.91$, $p < .001$, $\eta_p^2 = .268$, but perceived less selfish motives from low-status Distributors' strategic ($M = 3.74$, $SD = 1.87$) than open selfishness ($M = 3.03$, $SD = 2.06$), $F(1, 191) = 32.82$, $p < .001$, $\eta_p^2 = .147$. Other than that, neither status, $F(1, 191) = 0.129$, $p = .72$, $\eta_p^2 = .001$, nor behavior, $F(1, 191) = 3.37$, $p = .07$, $\eta_p^2 = .017$, had a significant

¹ Given the small ambiguity (i.e., a 1/36 chance to guess correctly) in the dice guessing procedure and the potential individual difference in perception of deception, we conducted parallel ANCOVAs controlling for perceived chances of winning and deception. These alternative analyses showed an identical pattern of results as reported here.

STATUS, EXPLOITATION, AND PUNISHMENT

main effect on perceived motives. A further within-participants mediation analysis revealed a significant mediation of motive perception (indirect effect = -0.08, 95% CI [-0.13, -0.04], $p < .001$), which, however, did not account for the status by behavior interaction effect on punishment severity (total effect = 0.06, 95% CI [-0.03, 0.14], $p = .19$; direct effect = 0.14, 95% CI [0.06, 0.22], $p = .001$).

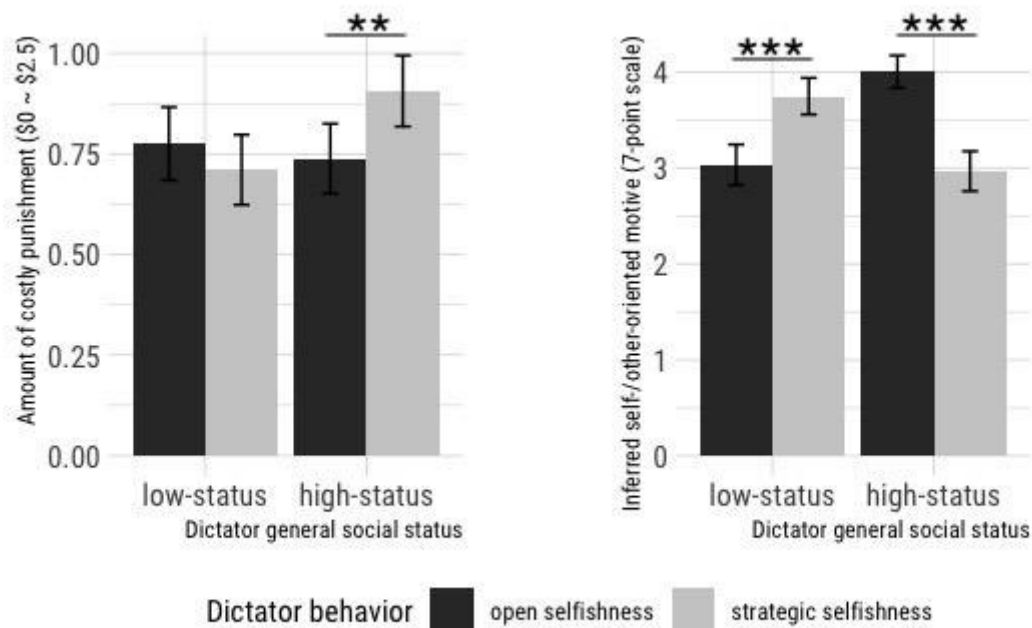


Figure 3. Observers' costly punishment (left panel) and inference of Dictator motives (right panel; with a lower score representing more self-interested motives) as a function of Dictator status and behavior in Study 2. ** $p < .01$. *** $p < .001$.

3.3 Discussion

Study 2 extended Study 1 by showing distinct observer punishment of high- versus average-status targets' strategic exploitation (as compared to open selfishness) in a dictator game. In addition to a selfish distribution, strategic communications (e.g., deceptive dice guessing) designed to earn an undeserved moral reputation incurred more punishment only

STATUS, EXPLOITATION, AND PUNISHMENT

when the targets possessed high social status. When the targets possessed a moderate social status, instead, people did not change their costly punishment depending on whether the selfish acts were accompanied by a strategic message or not. Moreover, people perceived strategic exploitation (vs. open selfishness) as particularly self-oriented when enacted by high-status instead of moderate-status targets.

4 General Discussion

From dyads to small groups, institutions, and societies, status asymmetry is often embedded in social interactions. Higher-status people, that is, those who are conferred more social prestige and esteem than others (Henrich & Gil-White, 2001; Magee & Galinsky, 2008), usually play a significant role in enforcing social norms. People rely on high- but not low-status people's norm communications to figure out what is the right thing to do, especially when the norms are ambivalent (e.g., Bauman et al., 2016; Henrich & Gil-White, 2001; Magee & Galinsky, 2008; Simpson et al., 2012). However, status holders are not always as trustworthy as they are expected to be; they can instead communicate strategically to maximize their self-interest. The current research demonstrates third-party punishment as a group mechanism to counteract status holders' misuse of peer trust and deference, such that observers are willing to incur relatively high personal costs to punish their strategic exploitation. Thus, the norm enforcement processes are not only about promoting competent and well-respected targets (Cheng et al., 2013; Henrich & Gil-White, 2001); it also demotes undeserved status holders by third-party sanctions.

Two pre-registered studies consistently revealed harsher third-party punishment of higher- (vs. lower-) status targets' strategic exploitation as compared to openly selfish transgressions. In a prisoner's dilemma (Study 1), people generated harsher punishment of higher-status defectors only when they had sent a misleading message advocating cooperation before their defection. In a dictator game (Study 2), selfishness incurred more

STATUS, EXPLOITATION, AND PUNISHMENT

severe punishment when disguised by an ostensibly fair procedure of dice guessing, and only for high- but not low-status holders. The punitive mechanism toward high-status strategic exploiters seemed to be determined by the deceptive nature of their transgressions but not the actual consequences, given that (1) the severity of punishment did not change depending on the victim loss (e.g., whether the partner cooperated [vs. defected] and thus lost more money; Study 1), and (2) observers also punished more harshly when status holders' exploitation only yielded reputational but not material benefits (Study 2).

4.1 Theoretical Implications

It has been questioned whether second-party sanctions can be an effective mechanism to deter status holders from exploitation (e.g., DeScioli & Kurzban, 2013; Risen & Gilovich, 2007). Motivated to augment their own survival and prosperity (Anderson et al., 2015), followers—whose self-interest is influenced by their relationship with leaders—can have blind trust in, and deference to, higher-status leaders. For example, followers evaluate ingroup transgressive leaders more favorably than other transgressive group members or outgroup leaders (Abrams et al., 2013). Followers can thus be less attentive to higher-status leaders' exploitative motives, interpret their transgressions favorably, and thus undermine the effectiveness of second-party sanctions. In contrast, the current research demonstrated that independent third-parties detected, and punished, status holders' efforts in exploiting others for their own benefits. Moreover, people deemed higher-status targets' strategic exploitation as more selfish than open transgressions (Studies 1 and 2). Status holders' strategic exploitation deviated more from cooperation and fairness norms and thus were punished more harshly. People's harsher punishment of higher-status strategic exploiters could have originated from their evolved instinct to maintain social norms (Carlsmith et al., 2002; Fehr & Fischbacher, 2004; Van Prooijen, 2018) but not necessarily evaluations of self-interest in potential interactions with the transgressors (Study 1). This is also consistent with Study 1

STATUS, EXPLOITATION, AND PUNISHMENT

findings of the status by Receiver choice interaction effect irrespective of the Sender's deception. People determined the validity of lower-status Sender's defection depending on partner choice (i.e., cooperation versus defection). In contrast, people may have expected higher-status Senders to better comply with cooperative norms, and thus punished their defection more harshly regardless of actual partner decisions.

The findings can also have broad implications on research on deception. Extending previous findings that people punish selfish actors more severely when they use deception (Boles et al., 2000; Brandts & Charness, 2003; Jordan et al., 2017; Ohtsubo et al., 2010), the current work revealed *when* people incur personal costs to sanction deception (e.g., the actors possess high social status). Levine and Schweitzer (2014, 2015) found that prosocial deceivers—who use deception to benefit others—are seen as more ethical and trustworthy than honest actors that do not benefit others. Corroborating this line of work, we highlighted the importance of motive perception in reactions to deception, such that people punished deceivers to the extent that they exploited others' trust and collective welfare. Such exploitation stimulated harsh punishment following both explicit deception (e.g., after directly stating that they would behave fairly; Study 2) and more implicitly (e.g., after indirectly implying that they would cooperate; Study 1).

4.2 Limitations and Future Directions

The current work went beyond moral judgment, and examined third-party behavioral punishment with actual costs, in the presence of status-based strategic exploitation. The findings were replicated across two studies with different norm violations (cooperation and fairness), and different kinds of exploitation (for material or reputational benefits). Some limitations should be noted, however. First, although we incentivized participants' punishment decisions with actual consequences on transgressors, only 5% of the participants received the actual bonus, which happened after they completed the experiments. Our

STATUS, EXPLOITATION, AND PUNISHMENT

findings cannot speak directly to perceived credibility or impact of the bonus, because our research did not examine whether participants believed that their decisions would have real impacts, or the extent to which they cared about the 5% chance to receive the actual bonus. Future research may replicate our findings in synchronized games, where third parties can make punitive decisions immediately after actual transgressive behaviors. Second, we observed a consistent pattern between perceived selfish motives and the severity of punishment of higher-status strategic exploiters. Despite so, mediation analyses did not support a mediation effect of motive perception in third-party punishment. Future research may consolidate our tentative findings and test more rigorously why people sacrifice own resources to punish higher-status strategic exploiters. For example, the punishment of higher-status exploiters can be a cost-effective way to signal punishers' virtuous qualities. Future studies may manipulate the anonymity of punishment, to examine whether costly punishment of status-based strategic exploitation occurs in anonymous settings (to maintain social norms), and whether people intensify their punishment in non-anonymous as compared to anonymous settings (to signal own virtues).

Moreover, the current research conceptualized status as mainly based on agentic leadership qualities like competence, while the implications of other status underpinnings like dominance and moral virtues on third-party punishment are not yet well-demonstrated (Bai et al., 2020). Different foundations of status can yield different behavior patterns (Cheng et al., 2013) and evoke different observer reactions to norm violations. Third, some scholars argue that laboratory findings of third-party punishment can be experimental artifacts (Pedersen et al., 2018). Future research may investigate punishment of status-based strategic exploitation in real-life situations (e.g., celebrities' environmental hypocrisy; Hofmann et al., 2018; Molho et al., 2020) and with an increased number of observer choices (punishing exploitation versus rewarding acts of integrity; Wang et al., 2009).

STATUS, EXPLOITATION, AND PUNISHMENT

5 Concluding Remarks

It is an inevitable fact from social life that groups often consist of people differing in status. This poses a threat to groups, in that members of high status may face opportunities for strategic exploitation. The current research adds credence to the idea that groups maintain cooperation, and discourage exploitation, by third-party punishment—a personally costly action that may often effectively counteract the tendencies of high-status members to exploit and misuse trust that is otherwise conferred with status. In other words, the present research illuminates the work on third-party punishment, to not simply “correct” norm violators but do so in a selective manner by focusing on members with high status whose actions would otherwise be most prone to undermine cooperation in groups.

ESM 1. Experimental materials.

This file details the experimental stimuli in both Studies 1 and 2.

STATUS, EXPLOITATION, AND PUNISHMENT

References

- Abrams, D., Randsley de Moura, G., & Travaglino, G. A. (2013). A double standard when group members behave badly: Transgression credit to ingroup leaders. *Journal of Personality and Social Psychology, 105*(5), 799-815. <https://doi.org/10.1037/a0033600>
- Anderson, C., Hildreth, J. A. D., & Howland, L. (2015). Is the desire for status a fundamental human motive? A review of the empirical literature. *Psychological Bulletin, 141*(3), 574–601. <https://doi.org/10.1037/a0038781>
- Bai, F., Ho, G. C. C., & Yan, J. (2020). Does virtue lead to status? Testing the moral virtue theory of status attainment. *Journal of Personality and Social Psychology, 118*(3), 501–531. <https://doi.org/10.1037/pspi0000192>
- Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin, 137*(4), 594–615. <https://doi.org/10.1037/a0023489>
- Barkow, J. H. (2014). Prestige and the ongoing process of culture revision. In J. T. Cheng, J. L. Tracy, & C. Anderson (Eds.), *The Psychology of Social Status* (p. 29–45). Springer, New York, NY. https://doi.org/10.1007/978-1-4939-0867-7_2
- Bauman, C. W., Tost, L. P., & Ong, M. (2016). Blame the shepherd not the sheep: Imitating higher-ranking transgressors mitigates punishment for unethical behavior. *Organizational Behavior and Human Decision Processes, 137*, 123-141. <https://doi.org/10.1016/j.obhdp.2016.08.006>
- Bøggild, T. (2020). Cheater detection in politics: Evolution and citizens' capacity to hold political leaders accountable. *The Leadership Quarterly, 31*(2), 101268. <https://doi.org/10.1016/j.leaqua.2018.09.006>

STATUS, EXPLOITATION, AND PUNISHMENT

- Boles, T. L., Croson, R. T., & Murnighan, J. K. (2000). Deception and retribution in repeated ultimatum bargaining. *Organizational Behavior and Human Decision Processes*, 83(2), 235-259. <https://doi.org/10.1006/obhd.2000.2908>
- Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, 100(6), 3531-3535. <https://doi.org/10.1073/pnas.0630443100>
- Brandts, J., & Charness, G. (2003). Truth or consequences: An experiment. *Management Science*, 49(1), 116-130. <https://doi.org/10.1287/mnsc.49.1.116.12755>
- Bunderson, J. S. (2003). Recognizing and utilizing expertise in work groups: A status characteristics perspective. *Administrative Science Quarterly*, 48(4), 557-591. <https://doi.org/10.2307/3556637>
- Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, 83(2), 284-299. <https://doi.org/10.1037/0022-3514.83.2.284>
- Case, C. R., Bae, K. K., & Maner, J. K. (2018). To Lead or to Be Liked: When Prestige-Oriented Leaders Prioritize Popularity Over Performance. *Journal of Personality and Social Psychology*, 115(4), 657-676. <http://dx.doi.org/10.1037/pspi0000138>
- Cheng, J. T., Tracy, J. L., Foulsham, T., Kingstone, A., & Henrich, J. (2013). Two ways to the top: Evidence that dominance and prestige are distinct yet viable avenues to social rank and influence. *Journal of Personality and Social Psychology*, 104(1), 103-125. <http://dx.doi.org/10.1037/a0030398>
- De Kwaadsteniet, E. W., & Van Dijk, E. (2010). Social status as a cue for tacit coordination. *Journal of Experimental Social Psychology*, 46(3), 515-524. <https://doi.org/10.1016/j.jesp.2010.01.005>

STATUS, EXPLOITATION, AND PUNISHMENT

- Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences*, *108*(32), 13335-13340. <https://doi.org/10.1073/pnas.1102131108>
- DeScioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin*, *139*(2), 477–496. <https://doi.org/10.1037/a0029065>
- DeScioli, P., & Bokemper, S. E. (2019). Intuitive Political Theory: People's Judgments About How Groups Should Decide. *Political Psychology*, *40*(3), 617-636. <https://doi.org/10.1111/pops.12528>
- Dong, M., van Prooijen, J. W., & van Lange, P. A. (2021). Calculating Hypocrites Effect: Moral judgments of word-deed contradictory transgressions depend on targets' competence. *Journal of Theoretical Social Psychology*. <https://doi.org/10.1002/jts5.113>
- Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63-87. [https://doi.org/10.1016/S1090-5138\(04\)00005-4](https://doi.org/10.1016/S1090-5138(04)00005-4)
- Hays, N. A., & Blader, S. L. (2017). To give or not to give? Interactive effects of status and legitimacy on generosity. *Journal of Personality and Social Psychology*, *112*(1), 17-38. <http://dx.doi.org/10.1037/pspi0000067>
- Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, *22*(3), 165-196. [https://doi.org/10.1016/S1090-5138\(00\)00071-4](https://doi.org/10.1016/S1090-5138(00)00071-4)
- Hofmann, W., Brandt, M. J., Wisneski, D. C., Rokenbach, B., & Skitka, L. J. (2018). Moral punishment in everyday life. *Personality and Social Psychology Bulletin*, *44*(12), 1697-1711. <https://doi.org/10.1177/0146167218775075>

STATUS, EXPLOITATION, AND PUNISHMENT

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473-476.

<https://doi.org/10.1038/nature16981>

Jordan, J. J., & Rand, D. G. (2020). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology*, *118*(1), 57-88.

<https://doi.org/10.1037/pspi0000186>

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological Science*, *28*(3), 356-368.

<https://doi.org/10.1177/0956797616685771>

Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit. *Psychological Science*, *27*(3), 405-418. <https://doi.org/10.1177/0956797615624469>

Levine, E. E., & Schweitzer, M. E. (2014). Are liars ethical? On the tension between benevolence and honesty. *Journal of Experimental Social Psychology*, *53*, 107-117.

<https://doi.org/10.1016/j.jesp.2014.03.005>

Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, *126*, 88-106.

<https://doi.org/10.1016/j.obhdp.2014.10.007>

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433-442. <https://doi.org/10.3758/s13428-016-0727-z>

Lönnqvist, J. E., Rilke, R. M., & Walkowitz, G. (2015). On why hypocrisy thrives: Reasonable doubt created by moral posturing can deter punishment. *Journal of*

STATUS, EXPLOITATION, AND PUNISHMENT

Experimental Social Psychology, 59, 139-145.

<https://doi.org/10.1016/j.jesp.2015.04.005>

Magee, J. C., & Galinsky, A. D. (2008). Social hierarchy: The self-reinforcing nature of power and status. *Academy of Management Annals*, 2(1), 351-398.

<https://doi.org/10.5465/19416520802211628>

McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition*, 134, 1-10. <https://doi.org/10.1016/j.cognition.2014.08.013>

Molho, C., Tybur, J. M., Van Lange, P. A., & Balliet, D. (2020). Direct and indirect punishment of norm violations in daily life. *Nature Communications*, 11(1), 1-9.

<https://doi.org/10.1038/s41467-020-17286-2>

Nelissen, R. M. (2008). The price you pay: Cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior*, 29(4), 242-248.

<https://doi.org/10.1016/j.evolhumbehav.2008.01.001>

Ohtsubo, Y., Masuda, F., Watanabe, E., & Masuchi, A. (2010). Dishonesty invites costly third-party punishment. *Evolution and Human Behavior*, 31(4), 259-264.

<https://doi.org/10.1016/j.evolhumbehav.2009.12.007>

Padilla, A., Hogan, R., & Kaiser, R. B. (2007). The toxic triangle: Destructive leaders, susceptible followers, and conducive environments. *The Leadership Quarterly*, 18(3), 176-194.

<https://doi.org/10.1016/j.leaqua.2007.03.001>

Pedersen, E. J., McAuliffe, W. H. B., & McCullough, M. E. (2018). The unresponsive avenger: More evidence that disinterested third parties do not punish

altruistically. *Journal of Experimental Psychology: General*, 147(4), 514-544.

<https://doi.org/10.1037/xge0000410>

STATUS, EXPLOITATION, AND PUNISHMENT

- Risen, J. L., & Gilovich, T. (2007). Target and observer differences in the acceptance of questionable apologies. *Journal of Personality and Social Psychology*, 92(3), 418–433. <https://doi.org/10.1037/0022-3514.92.3.418>
- Simons, T., Leroy, H., Collewaert, V., & Masschelein, S. (2015). How leader alignment of words and deeds affects followers: A meta-analysis of behavioral integrity research. *Journal of Business Ethics*, 132(4), 831-844. <http://dx.doi.org/10.1007/s10551-014-2332-3>
- Simpson, B., Willer, R., & Ridgeway, C. L. (2012). Status hierarchies and the organization of collective action. *Sociological Theory*, 30(3), 149-166. <https://doi.org/10.1177/0735275112457912>
- Van Prooijen, J. W. (2018). *The moral punishment instinct*. Oxford University Press.
- Wang, C. S., Galinsky, A. D., & Murnighan, J. K. (2009). Bad drives psychological reactions, but good propels behavior: Responses to honesty and deception. *Psychological Science*, 20(5), 634-644. <https://doi.org/10.1111/j.1467-9280.2009.02344.x>
- Willer, R. (2009). Groups reward individual sacrifice: The status solution to the collective action problem. *American Sociological Review*, 74(1), 23-43. <https://doi.org/10.1177/000312240907400102>
- Yudkin, D. A., Van Bavel, J. J., & Rhodes, M. (2020). Young children police group members at personal cost. *Journal of Experimental Psychology: General*, 149(1), 182–191. <https://doi.org/10.1037/xge0000613>