Author(s): Gordon, David; Lea, Stephen E. G.

Title: Who Punishes? : The Status of the Punishers Affects the Perceived Success of, and Indirect Benefits From, “Moralistic” Punishment

Year: 2016

Version:

Please cite the original version:

Gordon, D., & Lea, S. E. G. (2016). Who Punishes? : The Status of the Punishers Affects the Perceived Success of, and Indirect Benefits From, “Moralistic” Punishment. Evolutionary Psychology, 14(3). https://doi.org/10.1177/1474704916658042

Evolutionary Psychology
July-September 2016: 1–14

DOI: 10.1177/1474704916658042
evp.sagepub.com
SAGE

Who Punishes? The Status of the Punishers Affects the Perceived Success of, and Indirect Benefits From, "Moralistic" Punishment

David S. Gordon[1] and Stephen E. G. Lea[2]

Abstract

"Moralistic" punishment of free riders can provide a beneficial reputation, but the immediate behavior is costly to the punisher. In Study 1, we investigated whether variation in status would be perceived to offset or mitigate the costs of punishment. One hundred and nineteen participants were presented with a vignette describing a punishment scenario. Participants predicted whether punishment would occur, how successful it would be, and indicated their attitude to the punisher. Participants believed only intervention by a high-status (HS) individual would be successful and that low-status (LS) individuals would not intervene at all. HS individuals predicted to punish successfully were seen as more formidable and likable. Study 2 investigated whether punishment was necessary to maintain an HS position. One hundred and seventeen participants were presented with a vignette describing a punishment scenario. Participants were asked to indicate whether they wished to be led by the punisher. HS individuals who did not punish were less likely to be chosen as leaders compared to HS punishers, whereas LS individuals who punished were no more or less likely to be chosen than nonpunishers. The results of both studies suggest that only HS individuals are expected to punish, likely because such a position offsets some of the costs of punishment. As a result, only HS individual can access the reputation benefits from punishment. Furthermore, an HS position may be dependent on the willingness to punish antisocial behavior. The ramifications that these results may have for the evolution of moralistic punishment are discussed.

Keywords

punishment, status, fairness, third-party, leadership, reputation

Date received: January 10, 2016; Accepted: June 13, 2016

"Moralistic"[1] punishment, where an individual punishes the unfair, antisocial, or otherwise group detrimental behavior of another, has been shown to promote cooperative and prosocial behavior (Balliet, Mulder, & Van Lange, 2011; Fehr & Gächter, 2000). Punishment can promote such behavior even if it is delayed, whether in this life (Fudenberg & Pathak, 2010) or the next (McKay, Efferson, Whitehouse, & Fehr, 2010), and the mere presence of a third-party punisher can encourage prosocial activity (Halevy & Halali, 2015; Kim, Smith, & Brigham, 1998). While the group as a whole can benefit from the cooperative environment provided by punishment (e.g., Gächter, Renner, & Sefton, 2008), individuals who punish can be exploited by second-order free riders (Yamagishi, 1988), that is, group members who cooperate but do not pay the costs of punishment. Such exploitation, and other costs such as counter-punishment (Dreber & Rand, 2012), means the evolution of punishment as a mechanism to enforce cooperation remains difficult to explain (see West, Griffin, & Gardner, 2007).

Reputational Benefit of Punishment

This picture changes if punishers can gain from their actions. One mechanism is through reputation. Theoretical models demonstrate that reputational gains can allow punishment to evolve (Panchanathan & Boyd, 2004; Santos, Rankin, & Wedekind, 2011). Experimentally, punishers are found to be valued as social and sexual partners (Barclay, 2006; Farthing, 2005; Gordon, Madden, & Lea, 2014) and can be seen as trustworthy

[1] Department of Biological and Environmental Science, University of Jyväskylä, Jyväskylä, Finland
[2] University of Exeter, Exeter, UK

Corresponding Author:
David S. Gordon, Department of Biological and Environmental Science, University of Jyväskylä, PO Box 35, Jyväskylä 40014, Finland.
Email: david.s.gordon@jyu.fi

(Barclay, 2006; Jordan, Hoffman, Bloom, & Rand, 2016). Indeed, potential punishers seem to be sensitive to the presence of an audience (Bering, 2008; Kurzban, DeScioli, & O'Brien, 2007; but see Rockenbach & Milinski, 2011).

While the aforementioned suggests that punishers are *loved* by observers, this might not be the case. Reputation may be important for the evolution of punishment, but the reputation generated by punishment does not have to be an amiable one, that is, as a "nice" person (Gintis, Smith, & Bowles, 2001); punishers may instead be *feared*. For example, observers rate moralistic punishers to be equally as aggressive as individuals who engage in other, nonmoralistic, confrontational behavior (Gordon et al., 2014). While moralistic punishment is rarely characterized in the literature as an antagonistic act per se, any punishment—certainly in prestate societies (Mathew & Boyd, 2011) or informal settings (Kawakami, Dunn, Karmali, & Dovidio, 2009)—will inevitably involve individuals personally confronting the antisocial behavior of another. Any sort of confrontational behavior can act as a deterrent against future aggression from others (Benard, 2013) and individuals are less likely to cheat a punisher out of fear of retribution (Brandt, Hauert, & Sigmund, 2003). In fact, unless the motivations of the punisher are clear, observers are most likely to fear them (Raihani & Bshary, 2015).

Thus, on the one hand, punishers seem to be well liked, as they are trustworthy and can eliminate free riding. But, on the other, punishing shows personal formidability and indicates they should be treated with deference in any future interactions.

Barriers to Reputation Benefits

Regardless of any later returns from reputation, the immediate costs of punishment still represent a significant barrier. Firstly, experimentally punishment often needs to be cheap and effective, that is, for the ratio between the resources spent on punishment and those removed from the target to be low (for a review, see Balliet et al., 2011; see also, Egas & Riedl, 2008; Nikiforakis & Normann, 2008). While some punishment does occur at a high fee/fine ratio (Falk, Fehr, & Fischbacher, 2005), this ratio does not promote cooperation or deter free riding, as it is not seen as a deterrent (e.g., Markussen, Putterman, & Tyran, 2011; Nikiforakis & Normann, 2008). From an observer point-of-view, punishers who *fail*, insomuch as they do not alter the behavior of the target, are still well liked (Gordon et al., 2014), but given a choice, it is likely individuals would associate with punishers who could *actually* defend the public good (including themselves); both punishers may be interpersonally nice, but the latter is useful.

Secondly, perhaps the greatest cost to punishment is from retaliation/counterpunishment (Dreber & Rand, 2012), where the target of punishment responds in kind. When retaliation is possible in experiments, punishment is reduced (Cinyabuguma, Page, & Putterman, 2006; Nikiforakis, 2008). The threat of retaliation is a prime factor preventing otherwise cost-free punishment behavior, such as reporting criminal activity (Tarling & Morris, 2010) and might explain why direct moralistic punishment occurs far less frequently in nonstate compared to state societies (Hill, Barton, & Hurtado, 2009; Marlowe et al., 2008).

Simply put, any account of the reputational benefits generated by moralistic punishment must consider two factors. Firstly, conceptually, how a punisher can (a) cheaply inflict costs on the target and (b) survive, or at least believe they can withstand, any potential retaliation long enough to capitalize on those benefits. Secondly, and specifically for the current research, whether the expectations and opinions of observers are sensitive to these costs when making decisions *about* punishers.

Who Punishes?

Not all individuals experience the same costs of punishment, and such variation has implications for cooperation and punishment (Olson, 1965; for a recent review, see Singh & Boomsma, 2015). Theoretically, punishment could evolve if some individuals can punish more cheaply than others (de Weerd & Verbrugge, 2011). Experimentally, heterogeneity in the cost of punishment does induce cooperation (Bone & Raihani, 2015; Nikiforakis, Normann, & Wallace, 2009). It has been suggested that such heterogeneity can be the result of arbitrary proximate factors (Przepiorka & Diekmann, 2013). However, we suggest that status, or rather an individual's position in a social hierarchy,[2] might provide a consistent source of heterogeneity in the cost of punishment.

Firstly, individuals in a high-status (HS) position can punish more effectively, insomuch as they can inflict greater costs on the target physically (Sell, Tooby, & Cosmides, 2009) or use their position to limit or deny access to resources (Keltner, Gruenfeld, & Anderson, 2003; Maner & Mead, 2010). HS individuals also have a more extensive social network, and this can lower costs through the use of coalitional aggression and the ease with which it can be coordinated; in nonstate societies, the punishment of norm violations is coordinated and executed by individuals with strong coalition support (von Rueden, Gurven, Kaplan, & Stieglitz, 2014).

Secondly, we argue that HS individuals would also be at less risk from retaliation. *Dominance* is traditionally recognized by the fact an opponent yields without escalation (Drews, 1993), and humans will back down in the face of both formidable and prestigious opponents (Gambacorta & Ketelaar, 2013), will acquiesce to their demands (Nelissen & Meijers, 2011), and will otherwise avoid conflict with them (Jenson & Peterson, 2011). HS individual are expected to face a lower risk of retaliation after moralistically punishing (Gordon et al., 2014). Thus, while in principle the reputation benefits of punishment are open to all, only HS individuals are realistically able to access them.

Finally, reasoning about status hierarchies is a core part of human and nonhuman social cognition (Cummins, 2005; Thomsen, Frankenhuis, Ingold-Smith, & Carey, 2011), and an organism should, if possible, avoid conflicts that have a small likelihood of success (Maynard-Smith & Price, 1973). There are penalties for getting into such conflicts, for example, continued aggression from the victor (Clutton-Brock & Parker,

1995). Thus, as with any antagonistic encounter, we expect that moralistic punishment will be seen by observers in the context of status contests. In fact, if punishers are preferred because punishment signals prosocial qualities, then individuals should respond to consistency in the behavior (e.g., Számadó, 2011): All things being equal, status provides a *consistent* mechanism for lowering the cost of punishment, and observers should be sensitive to this when making judgments about punishers.

Current Studies: Status and Observer Opinions of Punishers

We argue that only HS individuals can realistically access the reputation benefits from punishment insomuch as low-status (LS) individuals are unlikely to intervene, or at least are not expected to. Study 1 investigated whether the reputation benefits of moralistic punishment are indeed confined to those in an HS position. Specifically, Study 1 investigated whether observers expect HS or LS individuals to punish, and what effect these expectations have on any reputation generated from punishment.

Study 2 investigated an additional aspect of the relationship between status and reputation. If, as we argue, HS is a prerequisite for punishment, then HS individuals potentially have access to an additional benefit, maintaining their HS position. Individuals prefer an environment where "someone" can punish (Gürerk, Irlenbusch, & Rockenbach, 2006) and will transfer power to individuals who are willing to punish noncooperation (Gross, Méder, Okamoto-Barth, & Riedl, 2016). Yet the benefits of HS, for example, a greater say in-group decision making, are often dependent on continuing to be a good social partner (von Rueden, Gurven, & Kaplan, 2008). Thus, any asymmetry in status may be accepted only as long as one is useful; punishing may be the *price* of occupying an HS position within a group. Study 2 specifically investigated the consequences that (not) punishing had on the reputation and status of HS/LS individuals.

Study 1

Gordon, Madden, and Lea (2014) found that punishment can be seen as a dominant act and that HS lowered the perceived risk of retaliation. Using the same experimental vignette method, Study 1 expands this past research in two key ways. Firstly, the current study allowed participants to make an active prediction about the outcome of any conflict based on the social status of the punisher. Secondly, it varied the status of the punisher and target. Doing both, these can potentially provide evidence that status is part of any judgments made about punishment and punishers. We predict that participants will only expect HS third parties to punish, although participants may expect third parties to punish less if the antisocial individual is also HS. We also predict that any reputational benefits will be a downstream result of the outcome participants predicted, that is, any effect that our manipulations of status may have on reputation will be mediated by how participants predicted the outcome of the scenario.

Method

Participants

Participants were recruited from the University of Exeter via a web-based recruitment system. A total of 119 participants, 26 males (M age = 24) and 93 females (M age = 20) with an overall age range of 18–46 completed the questionnaire. As an incentive, participants who completed the survey were entered into a prize draw for one of several £20 (US$36) store vouchers. No participants failed the manipulation checks (see *Manipulation checks and demographic questions*).

Material and Method

The survey was administered online. Participants followed an e-mail link, which randomly assigned them to one of the four conditions and were presented with a survey consisting of three sections. The first section presented participants with an experimental vignette and the second section collected participants' responses to the vignette. The third section collected demographic information and contained the manipulation check questions. The survey was presented to participants in the order shown later.

Experimental Vignettes

Participants were asked to imagine themselves as part of a local sports team, who, following an evening practice session, had retired to a local bar. The team had occupied a table, but there were not enough seats for everyone. Therefore, some members, including the participant, had to stand. Nearby, two strangers were sitting at another table and after a few minutes, one of them clearly headed to the bar to order drinks. Seeing this, one of the standing members of the team (the "transgressor") went over to the table and proceeded to take the now vacant chair, dismissing the objections of the still-seated stranger. When the transgressor returned with the chair, another member of the group (the "third party") was described as being visibly angered by this behavior. The scenario ended there without describing how this third party responded to the norm violation. Third party is used here to denote proximate disinterest insomuch as neither they nor an associate were harmed by the antisocial behavior.

The status of both the chair taker (the transgressor) and the other team member (the third party) was manipulated. Depending on the condition, each was described as either "a popular and skilled player" (HS) or "an unpopular and unskilled player" (LS), giving the study a 2×2 between-subjects design.

Social Perception Questions

Following the scenario, participants were asked to indicate "what happened next" from one of the three choices. They were asked to indicate whether they believed the third party would intervene successfully, with the transgressor returning the chair; the third party would intervene unsuccessfully, with the transgressor keeping the chair; or the third party would not

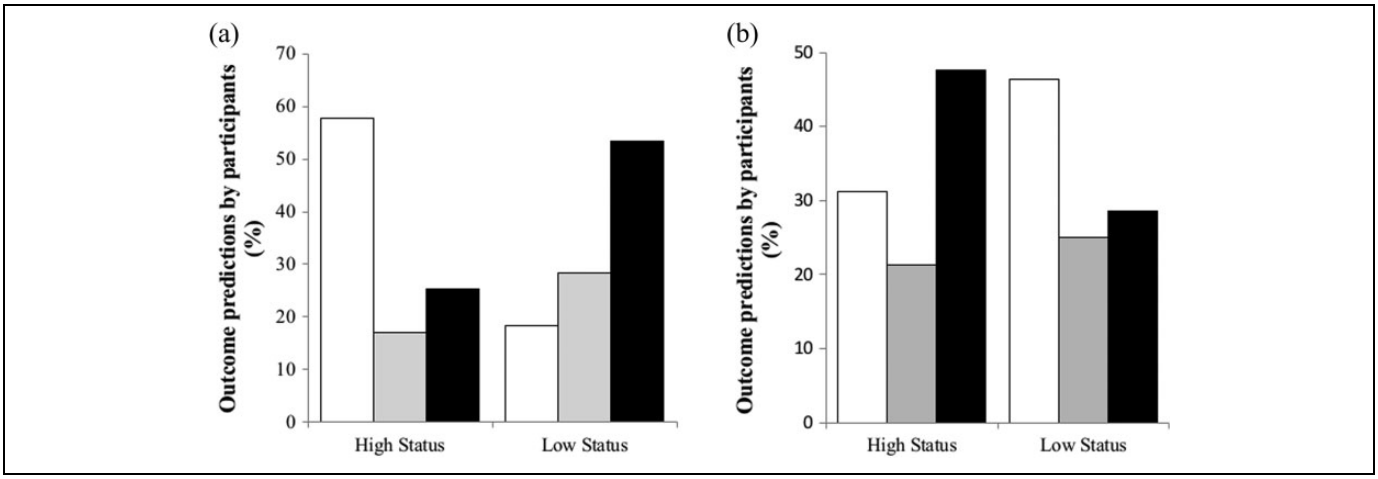

Figure 1. Predicted outcome of third-party punishment depending on (a) the rank of the third party or (b) the rank of the aggressor (white = successful intervention, gray = unsuccessful intervention, and black = no intervention).

intervene at all. The former two options stated that "after a brief exchange, the chair taker ... [did/did not return the chair]," that is to say it was not specified whether the intervention involved physical or social threats (see Gordon et al., 2014).

Participants were asked a series of questions regarding how likable the third party was. They were asked to rate the third party on a scale of 1 (*strongly disagree*) to 7 (*strongly agree*) as to how trustworthy, group focused, and "nice" they were, and whether they would work and socialize with the third party. These five questions had a high reliability index (Cronbach's $\alpha = .87$). Therefore, they were collapsed into a single "likability" variable for all future analyses.

Participants then answered a further set of questions concerning how dominant they perceived the third party to be. Participants rated on a scale of 1–7 (1 = *strongly disagree*, 7 = *strongly agree*), on how threatening, intimidating, dominant, antagonistic, or aggressive they perceived the third party to be. These five questions had a high reliability index ($\alpha = .86$) and were therefore collapsed into a single "dominance" variable for all future analyses. Finally, participants were asked to indicate, on a scale of 1 (*not likely at all*) to 7 (*extremely likely*), how likely it was that the transgressor would try and "get even" with the third party then or at a later date (retaliate).

Manipulation Checks and Demographic Questions

Participants were then asked the two comprehension questions. They were asked to indicate, from a choice of "popular and skilled," "unpopular and unskilled," or "sort of popular and skilled," how the transgressor and the third party were described in the scenario. Finally, participants indicated their age, sex, and nationality.

Statistical Analysis

The data were analyzed using SPSS 22. The outcome data were analyzed using a generalized linear model and all other data using analysis of variance (ANOVA). The mediation analyses were conducted using the PROCESS macro (Hayes, 2012).

Results

Outcome

Participants were first asked to indicate "what happened next": whether the third party successfully intervened, unsuccessfully intervened, or failed to intervene. As shown in Figure 1, participants believed that an HS third party would successfully intervene and that a subordinate third party was unlikely to intervene at all (Wald $\chi^2_1 = 18.33$, $p < .001$). As also shown in Figure 1, the rank of the transgressor also affected perceived outcome, with participants believing that a third party would be less likely to intervene when the transgressor was HS (Wald $\chi^2_1 = 5.03$, $p = .025$). Perceived outcome was not significantly affected by an interaction between the status of the third party and the transgressor (Wald $\chi^2_1 = 1.27$, $p = .26$). However, Figure 2 does suggest that while the status of the transgressor was important in the perceived outcomes, this was more the case when the third party was subordinate.

Likability

The status of the third party did not significantly affect their likability, $F(1, 115) = 2.57$, $p = .11$; however, the third party was less well liked when the transgressor was HS ($M = 4.6$, $SD = 1.2$) than when the transgressor was LS, $M = 5.0$, $SD = 0.9$; $F(1, 115) = 4.57$, $p = .035$). The likability of the third party was not significantly affected by an interaction between the status' of the third party and the transgressor, $F(1, 115) = 0.98$, $p = .75$).

A separate ANOVA was conducted using "outcome" as an independent variable. How participants predicted the outcome of the scenario had a strong effect on likability, $F(2, 116) = 4.11$, $p = .019$), with participants liking the third party who was predicted to be successful in their intervention ($M = 5.1$,

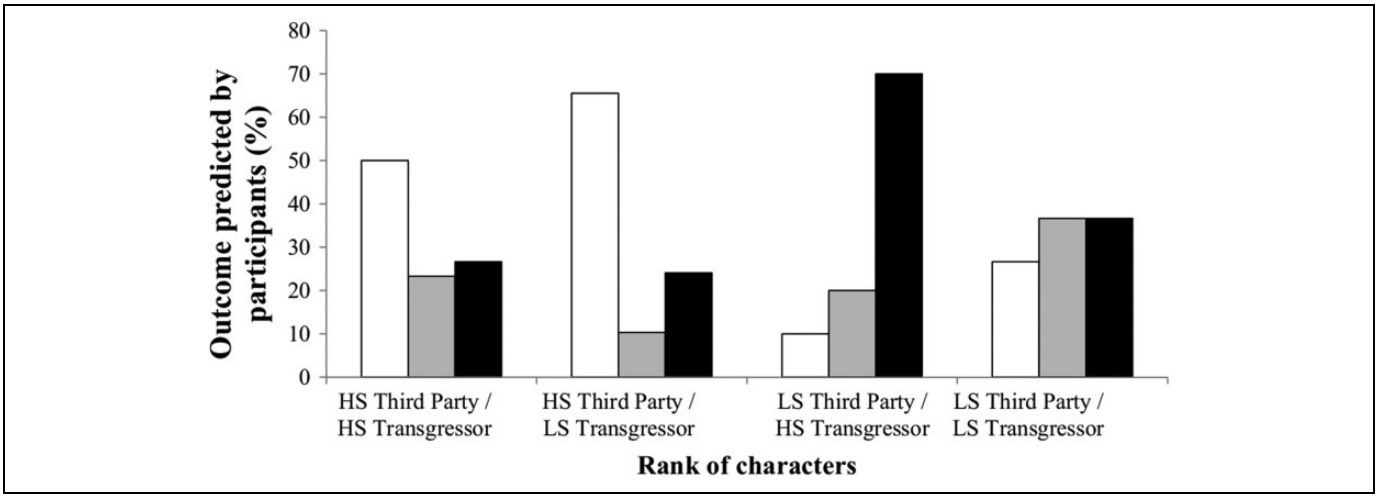

Figure 2. Predicted outcome of third-party punishment depending on the rank of the third party and the aggressor (white = successful intervention, gray = unsuccessful intervention, and black = no intervention).

$SD = 1.1$) more than those predicted to be unsuccessful ($M = 4.7$, $SD = 1.1$) or predicted to not intervene ($M = 4.5$, $SD = 0.9$). This might explain why participants liked the third party who punished an LS transgressor more, as successful punishment was seen to be less likely when directed against an HS transgressor (see Figure 1). Therefore, a mediation analysis was conducted with outcome as the mediating variable.[3] With the status of the third party controlled for, the predicted outcome completely mediated the relationship between the status of the transgressor and the likability of the third party ($b = 0.08$, BCa 95% CI [0.03, 0.24], on 5,000 samples), with the status of the transgressor no longer significantly affecting likability ($b = 0.34$, $t = 1.73$, $p = .09$). That is to say, the transgressor's status affected participant's likability ratings only insomuch as that status predicted the outcome of the scenario.

Dominance

As shown in Figure 3, unsurprisingly the third party was perceived to be more dominant when described as HS as opposed to subordinate, $F(1, 115) = 16.18$, $p < .001$. The third party was also marginally perceived to be more dominant when the transgressor they faced was described as LS, $F(1, 115) = 3.64$, $p = .059$; Figure 3. The perceived dominance of the third party was not significantly affected by an interaction between the status of the third party and the transgressor, $F(1, 115) = 0.24$, $p = .63$.

A separate ANOVA was conducted using outcome as an independent variable. As with likability, how participants predicted the outcome had a strong effect on perceived dominance, $F(2, 116) = 9.89$, $p < .001$, with successful third parties being seen as more dominant ($M = 3.8$, $SD = 1.2$) than unsuccessful ($M = 3.0$, $SD = 1.1$) or nonintervening ($M = 2.8$, $SD = 1.1$) third parties. With the status of the transgressor controlled for, the predicted outcome partially mediated the relationship between the status of the third party and their perceived dominance ($b = -0.23$, BCa 95% CI [−0.50, −0.08], on 5,000 samples), although the direct relationship between the two was still present ($b = -0.61$, $t = -2.75$, $p = .007$). That is to say, participant ratings of dominance were driven by both the outcome of the scenario, and whether the third party was described as HS or LS.

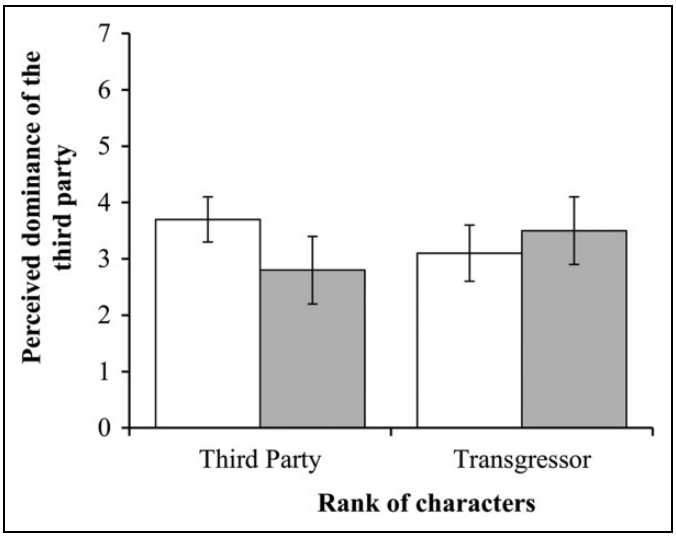

Figure 3. The perceived dominance of the third party depending on the rank of the third party and the aggressor (dominant = white and subordinate = gray). Error bars = 95% CI.

Interestingly, with the status of the third party controlled for, the predicted outcome of the interaction fully mediated the relationship between the status of the transgressor and the perceived dominance of the third party ($b = 0.11$, BCa 95% CI [0.01, 0.31], on 5,000 samples), with transgressor's status no longer significantly affecting dominance ($b = 0.29$, $t = 1.40$, $p = .17$); the effect of the transgressor's status on dominance ratings was entirely due to how this status affected the predicted outcome.

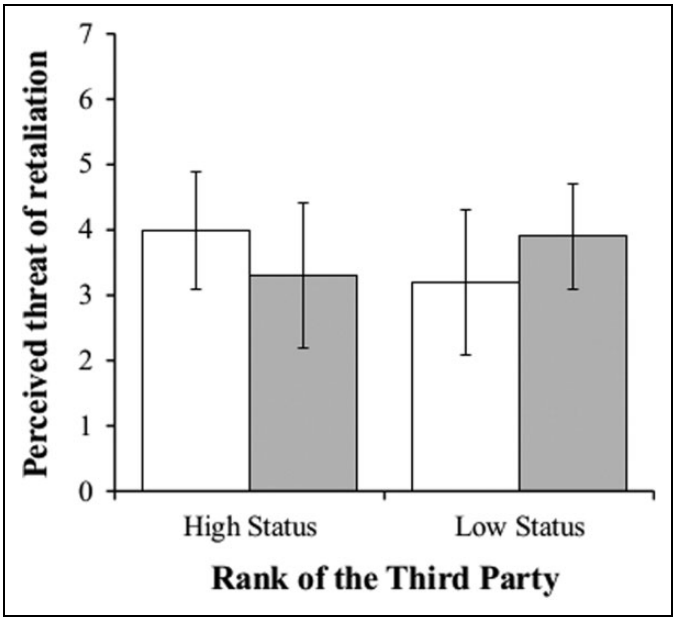

Figure 4. The perceived risk of retaliation against a successful or unsuccessful intervention depending on the rank of the third party and the aggressor (dominant aggressor = white and subordinate aggressor = gray). Error bars = 95% CI.

Retaliation

We assumed a priori that some participants would select the "do nothing" outcome. Accordingly, the retaliation item asked participants to "assume the agitated person [the third party] did intervene, regardless of your initial decision." Because data from these participants did not represent their true feelings, the analysis was run after removing participants who indicated that the third party would not intervene. As shown in Figure 4, while individually the status of the third party, $F(1, 67) = 0.005$, $p = .94$, and the transgressor, $F(1, 67) = 0.008$, $p = .93$, did not affect the perceived risk of retaliation, retaliation was affected by an interaction between the two, $F(1, 67) = 4.08$, $p = .047$); participants who predicted the third party would intervene felt that retaliation was more likely when the third party confronted an transgressor of equal status.

Discussion

Study 1 investigated whether the status of both an antisocial individual and a third party would affect how observers predicted the outcome of an act of moralistic punishment, and whether their relative status would have an impact on any subsequent reputational benefit. This was shown to be the case. The status of the belligerents was considered important by participants, with the rank of the third party influencing a "successful" outcome the most. While this is not surprising as such, it does support the suggestion made by Gordon et al. (2014) that subordinates might not be expected to intervene at all. Importantly, the pattern of predicted outcomes as shown in Figure 2 suggests that participants believed a punishing group member would either intervene successfully or not at all.

These results should be seen in the context of experimental punishment games, as in such games punishment is always *successful* insomuch as punish decisions always inflict costs on the target. We have suggested that only HS individuals are willing to punish because they can do so effectively due to physical formidability or social support; the fact that participants expected only HS individuals would confront antisocial behavior supports this conjecture.

Furthermore, any reputation, as either an amiable or intimidating individual (e.g., Barclay, 2006; Brandt et al., 2003; Raihani & Bshary, 2015), was dependent on a successful outcome for the punisher, which in turn was dependent on the status of the punisher and transgressor. While individuals prefer environments where punishment occurs (Gürerk et al., 2006), and thus might prefer punishers for the protection they afford, it is important that an individual can maintain their behavior (e.g., see dos Santos & Wedekind, 2015; Számadó, 2011). A subordinate may land, physically or metaphorically, a "lucky punch" but would be unlikely able to fend off the immediate retaliation or any subsequent feuds (e.g., Nikiforakis & Engelmann, 2011). Thus, there is a barrier to accessing the reputation benefits from punishment. As suggested by the results, this barrier can only be overcome by someone in an HS position.

Finally, there was some evidence that the punishment scenario was itself perceived in the context of a dominance/status contest. It was expected that retaliation risk would correspond to relative rank, that is, that a dominant punisher would face lower risk from a subordinate transgressor than a dominant one and that a subordinate punisher would face a greater risk from a dominant transgressor than a subordinate one. In fact, the risk of retaliation was perceived to be greater when the belligerents were of equal rank (Figure 4). This makes sense if participants perceived the encounter as a status contest rather than as a (purely) moralistic act as, within social hierarchies, conflict escalation should occur more between those of similar ranks (Stulp, Kordsmeyer, Buunk, & Verhulst, 2012; Wilson, 1980, pp. 141–142). Interestingly, the finding above was partly mirrored in a recent paper on punishment heterogeneity that found weak players were less likely to receive retaliation from stronger players (Bone, Wallace, Bshary, & Raihani, 2015), although here strong players were retaliated against by both player types. Additionally, while the framing of the belligerents as teammates could make retaliation unlikely, in real-life situations similar to the scenario, the opposite is true (Levine, Lowe, Best, & Heim, 2012). Thus, participant's belief about retaliation risk likely conforms to their real-life expectations experiences.

Study 2

Study 1 demonstrated that HS individuals can access the reputation benefits from punishment. However, any ambiguity in motive makes a punisher "feared" rather than "loved" (Raihani & Bshary, 2015), and nonpunishing cooperators are more

well liked than punishers (Jordan et al., 2016; Kiyonari & Barclay, 2008; Przepiorka & Liebe, 2015). Furthermore, whether any positive sentiment translates into physical gains is equivocal (Balafoutas, Nikiforakis, & Rockenbach, 2014; Nelissen, 2008). Thus, while Study 1 demonstrated that HS individuals are *expected* to punish, an important question is *why* they would be willing to punish?

One possible reason would be to maintain an HS position within a group. A HS position comes with intrinsic benefits (see Chapais, 2015; Rege, 2008), and while individuals prefer environments where punishment occurs (Gürerk et al., 2006), we dislike disadvantageous inequality (e.g., Leibbrandt & López-Pérez, 2011) and especially dislike individuals who become "too" domineering (Boehm & Boehm, 1999). Potentially, punishment may be the *price* of an HS position, that is, individuals are *allowed* to occupy a prominent position as long as some of their social power is used prosocially. In a number of nonstate societies, for example, a leadership position comes with the assumption that the leader will take part in dangerous activities (for a review, see Glowacki & von Rueden, 2015).

Study 2 was designed to test this suggestion by giving participants the option to remove the HS group member from a position of power. We predicted that HS individuals who *didn't* punish would lose their status. We also predicted that any decisions observers made about punishers would be mediated by how advantageous observers perceived an HS position to be.

Method

Participants

Participants were recruited from the University of Exeter (75) and the University of Dundee (42) via an e-mail advertisement sent to the undergraduate mailing lists. A total of 117 participants, 35 males (M age = 26), 82 females (M age = 23) with an overall age range of 18–52 completed the questionnaire. As an incentive, participants who completed the survey were entered into a prize draw for one of several £20 (US$32) store vouchers. There were no significant differences in the measured variables between the two institutions and, therefore, they were analyzed as one cohort. An additional 12 participants were excluded for failing at least one manipulation check (see *Manipulation checks and demographic questions*).

Materials and Procedure

The survey was administered online. Participants followed an e-mail link, which randomly assigned participants to one of the four conditions. They were then presented with a survey consisting of three sections. The first section presented participants with an experimental vignette and the second section collected participants' responses to the vignette. The third section collected demographic information and contained the manipulation check questions. The survey was presented to participants in the order shown later.

Experimental Vignettes

It was necessary to alter the scenario to a situation where the position of the HS individual was mutable (rather than a formal hierarchy such as one may find in an office) but was also "realistic" in the sense of Study 1 (no scenarios involving kings or revolutions). The scenario was therefore identical to that in Study 1 except for alterations in three areas. Firstly, participants were asked to imagine themselves as part of a university society, rather than a sports club. Secondly, the status of the third party was manipulated by describing them as either the current society president (HS) or a new member (LS). "New member" was used to suggest no group authority without providing personal information, such as "unpopular." Thirdly, any description of skill was removed as expertise is often valued over prosocial behavior (see von Rueden, Gavrilets, & Glowacki, 2015).

Finally, participants were informed whether the third party actually intervened or did nothing: given the results of Study 1, it was felt allowing participants to predict the outcome would yield an insufficient spread of responses to make comparisons between status and (lack of) intervention.

Social Perception Questions

Following the vignette, participants read that their society would be electing a new president soon. However, in order to run, a candidate had to be nominated anonymously by several society members first. Participants were asked how likely they would be to nominate the third party, on a scale of 1 (*definitely nominate someone else*) to 7 (*definitely nominate the current president/new member*).

Participants were then asked the same likability and dominance questions as in from Study 1. Both sets had a high reliability index (likable $\alpha = .89$; dominant $\alpha = .84$) and were collapsed into single "likability" and "dominant" variables for all future analyses.

Participants were then asked whether they believed the transgressor in the vignette would try and "get even" with the third party (retaliate), on a scale of 1 (*not at all likely*) to 7 (*very likely*).

We hypothesized that any relationship between status and punishment might be conditional on the benefit an HS individual extracts from that position. Therefore, participants were asked whether they believed the position of president was beneficial to the holder. These items were produced to reflect both the realities of a student society and the nonmaterial rewards leaders receive in nonstate societies (Glowacki & von Rueden, 2015). Such a focus can help explain why leadership might evolve in an environment where material rewards not are substantial or nonexistent. Participants were asked to respond, on a scale of 1 (*strongly disagree*) to 7 (*strongly agree*) to statements about the position of president that it (a) implicitly came with "perks," (b) was a reward in itself, (c) allowed one to help one's friends, (d) gave the holder access to opportunities not open to other members, and (e) whether having "the final say"

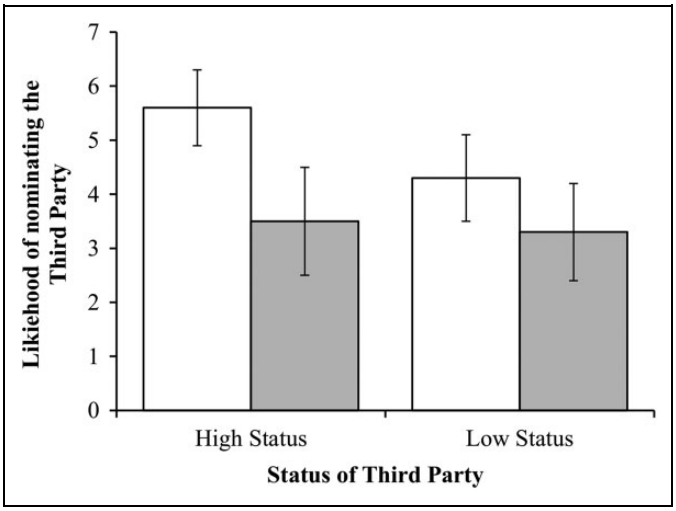

Figure 5. Likelihood of voting for the third party (punished = white and failed to punish = gray). Error bars = 95% CI.

on society issues was an advantage of the position. These five questions had a high reliability index ($\alpha = .78$) and were collapsed into a single "advantages" variable.

Manipulation Checks and Demographic Questions

Participants were then asked the two manipulation check questions. They were asked to indicate, from a choice of "president," "new member," or "stranger," how the third party was described in the scenario, and whether the third party "intervened" or "did nothing." Finally, participants indicated their age, sex, and nationality.

Statistical analysis. All data were analyzed using ANOVA in SPSS 22. The mediation analyses were conducted using the PROCESS macro (Hayes, 2012).

Results

Third-Party Punishment and Social Position

As shown by Figure 5, participants were far more willing to nominate the third party if they were labeled as HS, $F(1, 113) = 9.04$, $p = .003$, and, separately, if the third party intervened rather than did nothing, $F(1, 113) = 45.65$, $p < .001$. As also shown in Figure 5, willingness to nominate the third party was affected by an interaction between status and intervention, $F(1, 113) = 5.76$, $p = .018$; while nonintervention resulted in indifference to nominating the third party regardless of status, punishing HS third parties were far more likely to be nominated compared to LS punishers.

An analysis was conducted to investigate whether the "Advantages" ($M = 4.6$) variable mediated the relationship between status, intervention, and the willingness to nominate the third party. A mediation effect of Advantages did not occur ($b = -0.03$, $t = 0.24$, $p = .98$, 95% CI [−0.5, 0.4]), that is, belief about the advantages of the president's position did not affect participant responses to the manipulations.

Social responses to third party. As with Study 1, participants were also asked how likable and dominant they perceived the third party to be. As with Study 1, intervening third parties were seen as more likable, intervened, $M = 5.2$, $SD = 1.1$; did nothing, $M = 4.4$, $SD = 1.0$; $F(1, 113) = 16.50$, $p < .001$, but this perception was not affected by status, HS, $M = 4.9$, $SD = 1.1$; LS, $M = 4.7$, $SD = 1.0$; $F(1, 113) = 2.39$, $p = .13$, or any interaction of status and intervention, $F(1, 113) = 0.29$, $p = .59$. Likability correlated with nomination ($r = 0.32$, $n = 117$, $p < .001$) and partially mediated the relationship between the success of the third party and willingness to nominate ($b = -0.17$, BCa 95% CI [−0.40, −0.01], on 5,000 samples), although the direct relationship between the two was still present ($b = -1.26$, $t = -5.17$, $p < .001$), that is, the nomination results was not just an effect of participants "liking" the successful punishers.

Equally, intervening parties were seen as more dominant, intervened, $M = 2.9$, $SD = 1.1$; did nothing, $M = 2.4$, $SD = 1.2$; $F(1, 113) = 5.02$, $p = .027$, but this perception was not affected by status, HS, $M = 2.7$, $SD = 1.0$; LS, $M = 2.7$, $SD = 1.2$; $F(1, 113) = 0.18$, $p = .67$, or any interaction of the two, $F(1, 113) = 1.83$, $p = .18$. Dominance correlated with nomination ($r = .24$, $N = 117$, $p = .01$); however, dominance did not mediate the relationship between intervention and nomination ($b = -0.08$, BCa 95% CI [−0.27, −0.004], on 5,000 samples).

Thus, while the observers willingness to nominate the third party (confer/maintain status) was affected by an interaction between the latter's status and their actions, reputation was affected by the action of the third party alone.

Retaliation

The perceived threat of retaliation was not affected by the status of the third party, HS, $M = 3.7$, $SD = 1.5$; LS, $M = 3.9$, $SD = 1.6$; $F(1, 113) = 0.94$, $p = .33$, their intervention, intervened, $M = 3.6$, $SD = 1.6$; did nothing, $M = 4.1$, $SD = 1.5$; $F(1, 113) = 0.18$, $p = .08$, or an interaction between the two, $F(1, 113) = 0.29$, $p = .59$. Nor was the relationship between these factors and the perceived threat of retaliation mediated by any other variables.

Discussion

Study 2 investigated whether maintaining an HS position was dependent on moralistic punishment. On the core metric, willingness to nominate the third party, participants were more willing to vote for HS individuals who punished antisocial behavior. This, along with the likability and dominance ratings, suggests that moralistic punishment could be an effective mechanism to recruit and maintain social allies. While Study 1 and previous studies have shown that punishers are seen as dominant and likable, the current study suggests that the act of punishment can lead an individual to be given, or rather

allowed to keep, a formal position of leadership. Punishment is seen as the price for power.

Importantly, while participants were willing to nominate anyone who punished, the greatest difference in nominations was between HS punishers and nonpunishers (Figure 5). HS individuals are often expected to take on risky tasks as part of their position (von Rueden et al., 2015) and our result suggests that not only are HS individuals thought likely to punish (Study 1), but that their position becomes more precarious if they *fail* to act for the public good. Recently, it has been suggested that weaker individuals expect benevolent behavior from stronger individuals as a response to the fear of exploitation (Schilke, Reimann, & Cook, 2015). Instead, such expectations of prosociality could be seen as monitoring for behaviors that, if not conducted, would trigger a revolutionary coalition against a leader who attempts to behave too selfishly (Boehm & Boehm, 1999; see also, Van De Ven, Zeelenberg, & Pieters, 2010).

Thus, punishment by HS individuals may not be so much a case of the exploitation of the big by the small (Olson, 1965), but a trading of gains by the latter (cost-free cooperative environment) for gains by the former (the inherent advantages of HS within a social group). Indeed, the primary currency that leadership earns may be prestige (Price & Van Vugt, 2014), and our result suggests this can be taken away if a leader fails to moralistically punish.

We also hypothesized that the advantages that participants believed were part of an HS position would affect any reaction to (a lack of) moralistic punishment. This proved not to be the case. Two related possible explanations are as follows: participants did not value the advantages they believed the president of the society had or the advantages were not seen as being to the detriment of subordinates. Had the president received, for example, additional material advantages for their position (e.g., receiving a greater share of resources, which can provoke spiteful responses, Burns & Visser, 2006), or had participants observed one of the advantages in action (e.g., self-serving decision), then the advantages of an HS position might have produced an effect (e.g., Van Vugt, Jepson, Hart, & De Cremer, 2004).

Finally, status did not significantly affect the likelihood of retaliation. One explanation is the relative lack of social information provided by our scenario. Study 2 specifically chose to avoid explicit reference to physical or social attributes (e.g., sports skill[4] and popularity in Study 1). Physical attributes influence a variety of social decisions that, logically, they have no connection to in modern societies (von Rueden & van Vugt, 2015), and any confrontation of a norm violation at least has the potential to turn violent (see Levine, Taylor, & Best, 2011). Equally, coalitional support is also a key metric when the outcome of any contest is considered (e.g., Pun, Birch, & Baron, 2016). Nevertheless, our manipulation of status did imply social support. The variation in retaliation results between Studies 1, 2, and previous research (Gordon et al., 2014) could suggest that, if separated sufficiently, formidability and prestige may result in different expectations of punishers and punishment.

General Discussion

Study 1 demonstrated that any reputation generated from moralistic punishment is dependent on success, which itself is dependent on the status of the punisher. Study 2 found that punishment led to HS punishers maintaining their position and that a failure to punish led to this position being at risk. The willingness to grant punishers an HS position was also partially independent of how "liked" they were by participants. The results suggest that only HS individuals are expected to moralistically punish (Study1) and that the reputation benefits are strongly liked to status (Studies 1 and 2). Taken together, these results suggest the reputation of punishers is fundamentally tied to reasoning about social hierarchies.

Punishment and Reputation

In both studies, the reputation gained from punishment, as either a prosocial or intimidating individual (Barclay, 2006; Brandt et al., 2003; Raihani & Bshary, 2015), was dependent on a successful outcome of punishment, which in turn was dependent on the status of the punisher and transgressor (in Study 1). As stated previously, while individuals prefer environments where punishment occurs (Gürerk et al., 2006), it would be a mistake to join a group where the punishing individual could not act consistently in such a manner. In fact, the results of Study 2 suggest that the likelihood of *consistent* behavior is important, as HS individuals were more likely to be nominated when they successfully punished, compared to successful LS individuals. In Study 2, subordinate individuals were less likely to be nominated, even though successful intervention increased likability and perceived dominance across all status conditions (as in Study 1), and in Study 1, they were not expected to punish at all. A subordinate can still gain reputation from successfully punishing, yet the nomination result suggests participants were unwilling to grant LS punishers any authority. Thus, while punishment might not be a way to the top, it is a mechanism to remain there.

The fact that any reputation benefits afforded to punishers by observers were dependent on success, which was in turn dependent on status, should be stressed. By design, *all* punishment in economic experiments is successful, and this is not what individuals expect to occur in real social conflicts (see also Levine et al., 2011). While research has been concerned with downstream effects of successful punishment, on reputation (Nelissen, 2008), the behavior of free riders (Masclet, 2003), or group efficiency (Gächter et al., 2008), there has been no consideration that the attempt at punishment might fail and therefore of what determines success or failure, that is, who punishes *successfully*? The current studies demonstrate that, in the perceptions of observers at least, only punishment by HS individuals is likely to (a) succeed and (b) occur at all. Thus, past experiments are, in effect, investigating the behavior of individuals in an HS position, without recognizing that fact. While anger at an act of antisocial behavior might be ubiquitous, the ability to act upon it is realistically limited to powerful individuals.

Sex Effects

Our samples were heavily weighted toward females, but we made an a priori decision not to investigate sex differences. Firstly, a number of reviews (Cummins, 2005; Hawley, 2014; Hawley, Little, & Card, 2008) have suggested that sex differences in status seeking and contests have been overestimated. Males and females show similar behavior in conflict over resources (Griskevicius et al., 2009), in same-sex confrontations (Felson, 1982), and similar self-serving biases related to physical strength (Sell et al., 2009). Secondly, our studies concerned the *perception* of punishers rather than the act itself. Fundamentally, it is in the best interests of both males and females to (a) monitor the social environment and respond to events within it (Cummins, 1996, 1999) and (b) to recognize the cost/benefits of associating with certain individuals, for example, punishers.

This is not to say sex could not have had any effects: Females may prefer to establish status more covertly (Cummins, 2005), may value status differently to males depending on the circumstance (Snyder et al., 2011), and typical asymmetries in strength mean females may be less likely to *actually* punish outside of the laboratory (Levine et al., 2011; Parks, Osgood, Felson, Wells, & Graham, 2013). However, the investigation of such phenomena is beyond the aims of the current study. For the reasons mentioned earlier, we believe the female-biased participant pool did not affect the results of the current study or could detract from its conclusions.

Why Punish? Status and the Evolution of Punishment

Status influenced how observers perceive punishers and punishment. We suggest that status has a greater role in the evolution of punishment than just providing a proximate mechanism to overcome the immediate costs.

Firstly, status and social hierarchies are a core part of human social cognition (Hawley, 2014; Thomsen et al., 2011), and the need to out-maneuver one's rivals is a compelling explanation for the evolution of human intelligence (Byrne & Whiten, 1997; Dunbar, 1998; Jensen, 2010). It has been argued that dominance/status hierarchies, therefore, represent a set of basic implicit social norms (Cummins, 1996, 2005). As such, it is in the best interests of individuals, especially HS ones, to recognize when these rules are violated (Brosnan, 2011) and to punish others when violations occur (Clutton-Brock & Parker, 1995; Cummins, 1999). Indeed, we make "fair" decisions based on self-interest (DeScioli, Massenkoff, Shaw, Petersen, & Kurzban, 2014), and an HS position affects what is considered "fair" behavior (Pratto, Tatar, & Conway-Lanz, 1999; Sell et al., 2009). So in this regard alone, we should expect "moralistic" punishment and status to be closely associated.

Secondly, and more importantly, in human, and some nonhuman (De Waal, 1982/2007) societies, status is not just based on physical formidability but on social coalitions and political power (von Rueden et al., 2008). Coalitional aggression has a long history in human evolution (see Pietraszewski, Cosmides, & Tooby, 2014) and such aggression is often directed at individuals who are "too powerful" (Boehm & Boehm, 1999). Therefore, moralistic punishment provides a mechanism by which an individual can exert their status (Brandt et al., 2003; Gordon et al., 2014), but at the same time act as in a group beneficial way and thus be seen as "useful." Indeed, nonpunishers can benefit greatly from free riding on punishment (Roberts, 2013). This was highlighted by the nomination data, as this results was partially independent of likability; when deciding on a *leader*, competence—here the ability and willingness to punish—overrides any fear or envy observing punishment may generate.

Furthermore, while the current studies focused on heterogeneity in costs between individual punishers, punishment is often the result of coordinated group activity (see Guala, 2012). Nevertheless, individuals are needed to spearhead this coordination, and the attributes associated with an HS position (personal formidability, social support, etc.) would make coordination less risky and cheaper for such an individual (e.g., Boyd, Gintis, & Bowles, 2010). Participants, we expect, would likely react equally as negatively to an HS individual who did not coordinate punishment, as they did to one who did not punish. After all, such coordinating of group activity is expected of a leader (von Rueden et al., 2014), and leaders often pay immediate costs for this (Gavrilets, 2015). As stated earlier, punishment may be the price of an HS position.

The idea that punishment acts to *justify* an HS position is, at this point, entirely speculative. However, HS individuals certainly, both experimentally (Maner & Mead, 2010) and theoretically (Gavrilets & Fortunato, 2014; Powers & Lehmann, 2014), behave prosocially when it furthers their own ends.

Even without commenting on how status can affect direct benefit from enforcing cooperation (Raihani, Grutter, & Bshary, 2010; Singh & Boomsma, 2015; von Rueden & van Vugt, 2015), our status-based explanation for moralistic punishment provides the behavior with clear individual, condition-dependent, indirect benefits.

Conclusion

Recent studies on punishment have discursively acknowledged the role that interindividual differences in status might have in off-setting the costs of punishment. Specific emphasis has been placed on how the advantages of HS or dominant position relate to mechanisms in behavioral experiments, for example, the ability to punish effectively or retaliation risk (e.g., Bone et al., 2015; Roberts, 2013). This is in agreement with the anthropological literature where, whether formally recognized as "leaders" or not, it is HS individuals who tend to engage in punishment (for a review, see Glowacki & von Rueden, 2015). The current studies found that HS individuals were expected to punish (Study 1), HS individuals faced a greater risk of revolution should they fail to punish (Study 2), and the perceived risk of retaliation reflected status contests (Study 1). These results suggest that the perception of punishers and punishment is imbedded in the social cognition of status and reflects an

underlying strategy to remain in an HS position by demonstrating ''usefulness'' as a powerful yet benevolent ally: If the dilemma of leadership is how to gain (and maintain) ascendancy over others and simultaneously win their approval (Blau, 1964), our results suggest moralistic punishment is a good way for this to be achieved.

Human history, and indeed the human present, is filled with examples of individuals and groups claiming that their ability to ''protect us from threats'' justifies their position. Thus, our reasoning also has an intuitive logic about the function of, and motivation for, ''moralistically'' punishing antisocial behavior. We suggest that moralistic punishment evolved as a strategic behavior, in the context of greater coalitional aggression, to maintain status by demonstrating physical or social formidability, while at the same time acting in a ''pro social'' way that will not trigger group fission or revolution. We suggest that a focus on the status of punishers, for instance, by manipulating direct or indirect cues of status or leadership in experiments, would be a fruitful area of study. Equally, given the relationship between reputation and success, a more explicit study of whether a subordinate could in fact ''punish their way to the top'' would also expand on the ideas put forward in this article.

Acknowledgments

We would like to thank the two anonymous reviewers for their helpful comments on earlier versions of the article, and Jaime Benjamin for her assistance with the data collection for Study 2.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: We would like to acknowledge the financial support provided for this project by the College of Life and Environmental Sciences, University of Exeter.

Notes

1. While punishment in, for example, a public goods game (e.g., Fehr & Gächter, 2000) can be seen as different from punishment by a ''disinterested'' third party (e.g., Pedersen, Kurzban, & McCullough, 2013), both fundamentally describe the opportunistic—that is, not in response to direct antagonism—punishment of social norm violations. While ''costly punishment'' covers both these (see Guala, 2012), any realization of these costs may be conditional (Gordon et al., 2014) and given the vignette nature of the current studies, ''costly'' might be misleading. Thus, ''moralistic'' is used as a convenient short hand for any punishment of norm violations.
2. While formidability (individual fighting strength) and prestige (social regard) are conceptually different (see Henrich & Gil-White, 2001), in reality they can be hard to disentangle (Cheng, Tracy, Foulsham, Kingstone, & Henrich, 2013). For example, in nonstate societies, one impacts the other (e.g., von Rueden et al., 2008) and in modern societies, they cause similar behavioral effects from others (e.g., Gambacorta & Ketelaar, 2013). For the sake of clarity and simplicity, and because our aim is not to tackle this issue per se, we will use ''status'' as a label, as it reflects our concept across various fields of biology and psychology (see Cheng et al., 2013, table 1).
3. Analyses carried out using linear regressions suggest that the ''outcome'' categories produce a graded response and can, therefore, be considered as a ''scale of intervention,'' from likelihood of no intervention to certain success.
4. We would argue that being labeled as ''good at sport'' implies one is physically fit/strong.

References

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences*, *111*, 15924–15927.

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, *137*, 594.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*, 325–344. doi:10.1016/j.evolhumbehav.2006.01.003

Benard, S. (2013). Reputation systems, aggression, and deterrence in social interaction. *Social Science Research*, *42*, 230–245.

Bering, J. (2008). The effects of perceived anonymity on altruistic punishment. *Evolutionary Psychology*, *6*, 487–501.

Blau, P. M. (1964). *Exchange and power in social life*. Piscataway, NJ: Transaction Publishers.

Boehm, C., & Boehm, C. (1999). *Hierarchy in the forest: The evolution of egalitarian behavior*. Cambridge, MA: Harvard University Press.

Bone, J. E., & Raihani, N. J. (2015). Human punishment is motivated by both a desire for revenge and a desire for equality. *Evolution and Human Behavior*, *36*, 323–330.

Bone, J. E., Wallace, B., Bshary, R., & Raihani, N. J. (2015). The effect of power asymmetries on cooperation and punishment in a prisoner's dilemma game. *PLoS One, 10*, e0117183.

Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*, *328*, 617–620. doi:10.1126/science.1183665

Brandt, H., Hauert, C., & Sigmund, K. (2003). Punishment and reputation in spatial public goods games. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*, 1099–1104.

Brosnan, S. F. (2011). An evolutionary perspective on morality. *Journal of Economic Behavior & Organization*, *77*, 23–30.

Burns, J., & Visser, M. (2006). *Bridging the great divide in South Africa: Inequality and punishment in the provision of public goods* (Rapport nr.: Working Papers in Economics 219). Göteborg, Sweden: Göteborg University.

Byrne, R. W., & Whiten, A. (1997). Machiavellian intelligence. In A. Whiten & R. W. Byrne (Eds.), *Machiavellian intelligence II: Extensions and evaluations* (pp. 1–23). Cambridge, MA: Cambridge University Press.

Chapais, B. (2015). Competence and the evolutionary origins of status and power in humans. *Human Nature*, *7*. doi:10.1007/s12110-015-9227-6

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*, 103.

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

Clutton-Brock, T., & Parker, G. (1995). Punishment in animal societies. *Nature*, *373*, 209–216.

Cummins, D. (1996). Dominance hierarchies and the evolution of human reasoning. *Minds and Machines*, *6*, 463–480.

Cummins, D. (1999). Cheater detection is modified by social rank: The impact of dominance on the evolution of cognitive functions. *Evolution and Human Behavior*, *20*, 229–248.

Cummins, D. (2005). Dominance, status, and social hierarchies. In D. Buss (Ed.), *The handbook of evolutionary psychology* (pp. 676–697). Hoboken, NJ: John Wiley.

De Waal, F. (1982/2007). *Chimpanzee politics: Power and sex among apes* (25th anniversary ed.). Baltimore, MD: Johns Hopkins University Press.

de Weerd, H., & Verbrugge, R. (2011). Evolution of altruistic punishment in heterogeneous populations. *Journal of Theoretical Biology*, *290*, 88–103.

DeScioli, P., Massenkoff, M., Shaw, A., Petersen, M. B., & Kurzban, R. (2014). Equity or equality? Moral judgments follow the money. *Proceedings of the Royal Society B: Biological Sciences*, *281*, 2014–2112.

dos Santos, M., & Wedekind, C. (2015). Reputation based on punishment rather than generosity allows for evolution of cooperation in sizable groups. *Evolution and Human Behavior*, *36*, 59–64.

Dreber, A., & Rand, D. G. (2012). Retaliation and antisocial punishment are overlooked in many theoretical models as well as behavioral experiments. *The Behavioral and Brain Sciences*, *35*, 24.

Drews, C. (1993). The concept and definition of dominance in animal behaviour. *Behaviour*, *125*, 283–313.

Dunbar, R. I. M. (1998). The social brain hypothesis. *Brain*, *9*, 10.

Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences*, *275*, 871–878. doi:10.1098/rspb.2007.1558

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, *73*, 2017–2030. doi:10.1111/j.1468-0262.2005.00644.x

Farthing, G. W. (2005). Attitudes toward heroic and nonheroic physical risk takers as mates and as friends. *Evolution and Human Behavior*, *26*, 171–185.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, *90*, 980–994.

Felson, R. B. (1982). Impression management and the escalation of aggression and violence. *Social Psychology Quarterly*, *45*, 245–254.

Fudenberg, D., & Pathak, P. A. (2010). Unobserved punishment supports cooperation. *Journal of Public Economics*, *94*, 78–86. doi:10.1016/j.jpubeco.2009.10.007

Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science*, *322*, 1510.

Gambacorta, D., & Ketelaar, T. (2013). Dominance and deference: Men inhibit creative displays during mate competition when their competitor is strong. *Evolution and Human Behavior*, *34*, 330–333.

Gavrilets, S. (2015). Collective action problem in heterogeneous groups. *Philosophical Transactions of the Royal Society B*, *370*, 20150016.

Gavrilets, S., & Fortunato, L. (2014). A solution to the collective action problem in between-group conflict with within-group inequality. *Nature Communications*, *5*, 3526.

Gintis, H., Smith, E., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of Theoretical Biology*, *213*, 103–119.

Glowacki, L., & von Rueden, C. R. (2015). Leadership solves collective action problems in small-scale societies. *Philosophical Transactions of the Royal Society B*, *370*, 20150010.

Gordon, D. S., Madden, J. R., & Lea, S. E. G. (2014). Both loved and feared: Third party punishers are viewed as formidable and likeable, but these reputational benefits may only be open to dominant individuals. *PLoS One*, *9*, e110045. doi:10.1371/journal.pone.0110045

Griskevicius, V., Tybur, J. M., Gangestad, S. W., Perea, E. F., Shapiro, J. R., & Kenrick, D. T. (2009). Aggress to impress: Hostility as an evolved context-dependent strategy. *Journal of Personality and Social Psychology*, *96*, 980.

Gross, J., Méder, Z. Z., Okamoto-Barth, S., & Riedl, A. (2016). Building the Leviathan–Voluntary centralisation of punishment power sustains cooperation in humans. *Scientific Reports*, *6*, Article no. 20767.

Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, *35*, 1.

Gürerk, O., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, *312*, 108–111 doi:10.1126/science.1123633

Halevy, N., & Halali, E. (2015). Selfish third parties act as peacemakers by transforming conflicts and promoting cooperation. *Proceedings of the National Academy of Sciences*, *112*, 6937–6942.

Hawley, P. H. (2014). Ontogeny and social dominance: A developmental view of human power patterns. *Evolutionary Psychology*, *12*, 318–342.

Hawley, P. H., Little, T. D., & Card, N. A. (2008). The myth of the alpha male: A new look at dominance-related beliefs and behaviors among adolescent males and females. *International Journal of Behavioral Development*, *32*, 76.

Hayes, A. F. (2012). *PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling*. Unpublished Manuscript, University of Kansas, KS.

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*, 165–196.

Hill, K., Barton, M., & Hurtado, A. M. (2009). The emergence of human uniqueness: Characters underlying behavioral modernity. *Evolutionary Anthropology: Issues, News, and Reviews*, *18*, 187–200.

Jensen, K. (2010). Punishment and spite, the dark side of cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*, 2635–2650. doi:10.1098/rstb.2010.0146

Jenson, N. H., & Peterson, M. B. (2011). To defer or to stand up? How offender formidability affects third party moral outrage. *Evolutionary Psychology*, *9*, 118–136.

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

Kawakami, K., Dunn, E., Karmali, F., & Dovidio, J. F. (2009). Mispredicting affective and behavioral responses to racism. *Science*, *323*, 276–278.

Keltner, D., Gruenfeld, D. H., & Anderson, C. (2003). Power, approach, and inhibition. *Psychological Review*, *110*, 265.

Kim, S. H., Smith, R. H., & Brigham, N. L. (1998). Effects of power imbalance and the presence of third parties on reactions to harm: Upward and downward revenge. *Personality and Social Psychology Bulletin*, *24*, 353.

Kiyonari, T., & Barclay, P. (2008). Cooperation in social dilemmas: Free riding may be thwarted by second-order reward rather than by punishment. *Journal of Personality and Social Psychology*, *95*, 826.

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*, 75–84.

Leibbrandt, A., & López-Pérez, R. (2011). The dark side of altruistic third-party punishment. *Journal of Conflict Resolution*, *55*, 761–784.

Levine, M., Lowe, R., Best, R., & Heim, D. (2012). 'We police it ourselves': Group processes in the escalation and regulation of violence in the night-time economy. *European Journal of Social Psychology*, *42*, 924–932.

Levine, M., Taylor, P. J., & Best, R. (2011). Third parties, violence, and conflict resolution: The role of group size and collective action in the microregulation of violence. *Psychological Science*, *22*, 406–412.

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. *Journal of Personality and Social Psychology*, *99*, 482.

Markussen, T., Putterman, L., & Tyran, J.-R. (2011). *Self-organization for collective action: An experimental study of voting on formal, informal, and no sanction regimes* (Working Paper 2011–4). Providence, RI: Department of Economics, Brown University.

Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., . . . Tracer, D. (2008). More 'altruistic' punishment in larger societies. *Proceedings of the Royal Society B: Biological Sciences*, *275*, 587–592. doi:10.1098/rspb.2007.1517

Masclet, D. (2003). Ostracism in work teams: A public good experiment. *International Journal of Manpower*, *24*, 867–887.

Mathew, S., & Boyd, R. (2011). Punishment sustains large-scale cooperation in pre-state warfare. *Proceedings of the National Academy of Sciences*. doi:PNAS 2011: 1105604108v1-201105604

Maynard-Smith, J., & Price, G. (1973). The logic of animal conflict. *Nature*, *246*, 15–18. doi:10.1038/246015a0

McKay, R., Efferson, C., Whitehouse, H., & Fehr, E. (2010). Wrath of God: Religious primes and punishment. *Proceedings of the Royal Society B: Biological Sciences*, *278*, 1858–1863. doi:10.1098/rspb.2010.2125

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

Nelissen, R., & Meijers, M. H. (2011). Social benefits of luxury brands as costly signals of wealth and status. *Evolution and Human Behavior*, *32*, 343–355.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics*, *92*, 91–112. doi:10.1016/j.jpubeco.2007.04.008

Nikiforakis, N., & Engelmann, D. (2011). Altruistic punishment and the threat of feuds. *Journal of Economic Behavior & Organization*, *78*, 319–332.

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

Nikiforakis, N., Normann, H., & Wallace, B. (2009). Asymmetric enforcement of cooperation in a social dilemma. *Southern Economic Journal*, *76*, 638–659.

Olson, M. (1965). *Logic of collective action public goods and the theory of groups* (Rev. ed.). Cambridge, MA: Harvard University Press.

Panchanathan, K., & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, *432*, 499–502.

Parks, M. J., Osgood, D. W., Felson, R. B., Wells, S., & Graham, K. (2013). Third party involvement in barroom conflicts. *Aggressive Behavior*, *39*, 257–268.

Pedersen, E. J., Kurzban, R., & McCullough, M. E. (2013). Do humans really punish altruistically? A closer look. *Proceedings of the Royal Society B: Biological Sciences*, *280*, 1–8.

Pietraszewski, D., Cosmides, L., & Tooby, J. (2014). The content of our cooperation, not the color of our skin: An alliance detection system regulates categorization by coalition and race, but not sex. *PLoS One*, *9*, e88534.

Powers, S. T., & Lehmann, L. (2014). An evolutionary model explaining the Neolithic transition from egalitarianism to leadership and despotism. *Proceedings of the Royal Society B: Biological Sciences*, *281*, 20141349.

Pratto, F., Tatar, D. G., & Conway-Lanz, S. (1999). Who gets what and why: Determinants of social allocations. *Political Psychology*, *20*, 127–150.

Price, M. E., & Van Vugt, M. (2014). The evolution of leader–follower reciprocity: The theory of service-for-prestige. *Frontiers in Human Neuroscience*, *8*.

Przepiorka, W., & Diekmann, A. (2013). Individual heterogeneity and costly punishment: A volunteer's dilemma. *Proceedings of the Royal Society B: Biological Sciences*, *280*, 2013–2247.

Przepiorka, W., & Liebe, U. (2015). Generosity is a sign of trustworthiness–The punishment of selfishness is not. *Evolution and Human Behavior*. doi:10.1016/j.evolhumbehav.2015.12.003

Pun, A., Birch, S. A., & Baron, A. S. (2016). Infants use relative numerical group size to infer social dominance. *Proceedings of the National Academy of Sciences*, *113*, 2376–2381.

Raihani, N. J., & Bshary, R. (2015). The reputation of punishers. *Trends in Ecology & Evolution*, *30*, 98–103.

Raihani, N. J., Grutter, A. S., & Bshary, R. (2010). Punishers benefit from third-party punishment in fish. *Science*, *327*, 171.

Rege, M. (2008). Why do people care about social status? *Journal of Economic Behavior & Organization*, *66*, 233–242.

Roberts, G. (2013). When punishment pays. *PLoS One*, *8*, e57378.

Rockenbach, B., & Milinski, M. (2011). To qualify as a social partner, humans hide severe punishment, although their observed cooperativeness is decisive. *Proceedings of the National Academy of Sciences*, *108*, 18307–18312.

Santos, M. D., Rankin, D. J., & Wedekind, C. (2011). The evolution of punishment through reputation. *Proceedings of the Royal Society B: Biological Sciences*, *278*, 371–377.

Schilke, O., Reimann, M., & Cook, K. S. (2015). Power decreases trust in social exchange. *Proceedings of the National Academy of Sciences*, *112*, 12950–12955.

Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences*, *106*, 15073–15078.

Singh, M., & Boomsma, J. J. (2015). Policing and punishment across the domains of social evolution. *Oikos*, *124*, 971–982.

Snyder, J. K., Fessler, D. M., Tiokhin, L., Frederick, D. A., Lee, S. W., & Navarrete, C. D. (2011). Trade-offs in a dangerous world: Women's fear of crime predicts preferences for aggressive and formidable mates. *Evolution and Human Behavior*, *32*, 127–137.

Stulp, G., Kordsmeyer, T., Buunk, A. P., & Verhulst, S. (2012). Increased aggression during human group contests when competitive ability is more similar. *Biology Letters*, *8*, 921–923.

Számadó, S. (2011). Long-term commitment promotes honest status signalling. *Animal Behaviour*, *82*, 295–302.

Tarling, R., & Morris, K. (2010). Reporting crime to the police. *British Journal of Criminology*, *50*, 474.

Thomsen, L., Frankenhuis, W. E., Ingold-Smith, M., & Carey, S. (2011). Big and mighty: Preverbal infants mentally represent social dominance. *Science*, *331*, 477–480.

Van De Ven, N., Zeelenberg, M., & Pieters, R. (2010). Warding off the evil eye: When the fear of being envied increases prosocial behavior. *Psychological Science*, *21*, 1671–1677.

Van Vugt, M., Jepson, S. F., Hart, C. M., & De Cremer, D. (2004). Autocratic leadership in social dilemmas: A threat to group stability. *Journal of Experimental Social Psychology*, *40*, 1–13.

von Rueden, C. R., Gavrilets, S., & Glowacki, L. (2015). Solving the puzzle of collective action through inter-individual differences. *Philosophical Transactions of the Royal Society B*, *370*, 20150002.

von Rueden, C. R., Gurven, M., & Kaplan, H. (2008). The multiple dimensions of male social status in an Amazonian society. *Evolution and Human Behavior*, *29*, 402–415.

von Rueden, C. R., Gurven, M., Kaplan, H., & Stieglitz, J. (2014). Leadership in an egalitarian society. *Human Nature*, *25*, 538–566.

von Rueden, C. R., & van Vugt, M. (2015). Leadership in small-scale societies: Some implications for theory, research, and practice. *The Leadership Quarterly*, *26*, 978–990.

West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology*, *20*, 415–432.

Wilson, E. O. (1980). *Sociobiology: The abridged version*. Cambridge, MA: Harvard University Press.

Yamagishi, T. (1988). The provision of a sanctioning system in the United States and Japan. *Social Psychology Quarterly*, *51*, 265–271.