Direct Punishment and Indirect Reputation-Based Tactics to Intervene Against Offenses

Catherine Molho[1,2] & Junhui Wu[3,4]

The final version of this paper is published in *Philosophical Transactions of the Royal Society B*, 376(1838), 20200289. https://doi.org/10.1098/rstb.2020.0289.

[1] *Institute for Advanced Study in Toulouse, Université Toulouse 1 Capitole*

[2] *Center for Research in Experimental Economics and Political Decision Making (CREED), University of Amsterdam*

[3] *CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences*

[4] *Department of Psychology, University of Chinese Academy of Sciences*

* Corresponding author | c.molho@uva.nl

Center for Research in Experimental Economics and Political Decision Making (CREED)

University of Amsterdam

Roetersstraat 11

1001 NJ Amsterdam, The Netherlands

Abstract

Punishment and reputation-based mechanisms play a major role in supporting the evolution of human cooperation. Theoretical accounts and field observations suggest that humans use multiple tactics to intervene against offenses—including confrontation, gossip, and ostracism—which have unique benefits and costs. Here, we draw a distinction between *direct* punishment tactics (i.e., physical and verbal confrontation) and *indirect* reputation-based tactics (i.e., gossip and ostracism). Based on this distinction, we sketch the common and unique social functions that different tactics are tailored to serve and describe information-processing mechanisms that potentially underlie decisions concerning how to intervene against offenses. We propose that decision rules guiding direct and indirect tactics should weigh information about the benefits of changing others' behavior versus the costs of potential retaliation. Based on a synthesis of existing evidence, we highlight the role of situational, relational, and emotional factors in motivating distinct punishment tactics. We suggest that delineating between direct and indirect tactics can inform debates about the prevalence and functions of punishment, and the reputational consequences of third-party intervention against offenses. We emphasize the need to study how people use reputation-based tactics for partner recalibration and partner choice, within interdependent relationships and social networks, and in daily life situations.

Keywords: *punishment, reputation, gossip, partner choice, ostracism, cooperation*

1. Introduction

Across societies, individuals and communities face challenges in terms of maintaining cooperation, deterring free-riding on public goods, and ensuring adherence to social norms (1–3). Theoretical models and experiments have shown that punishment via the selective imposition of costs on non-cooperators and norm violators can support the evolution of human cooperation (4–8). In experimental settings, individuals punish offenders even at a personal cost, though there is substantial cross-cultural variation in punishment norms (2,9,10). That said, multiple complementary mechanisms have been proposed to explain the evolution of cooperation, including reputation-based indirect reciprocity (11,12) and partner choice (13–15). Experiments have provided evidence for these mechanisms in action showing that gossip and ostracism can promote cooperation (16–20), perhaps more efficiently than punishment (21,22; cf. 23,24).

This review aims to contribute to understanding the unique antecedents and consequences of the various punishment and reputation-based tactics that humans use to intervene against non-cooperators and norm transgressors. Based on prior work (5,25,26), we define punishment as a response to an offense via inflicting some costs on the offender[1]. While punishment might be aimed at changing an offender's (future) behavior, we do not consider deterrence as a necessary component of its definition. For example, punishers can aim at reducing disadvantageous inequality or creating advantageous inequality without deterrence (30,31), and they can reap reputational benefits independent of any recalibration of offenders' behavior. Moreover, we use an inclusive definition of punishment that considers a host of tactics used to inflict costs on

[1] When describing the functions and mechanisms underlying direct and indirect tactics, we treat punishment as having potential long-term benefits for punishers. Whether and how punishment that involves fitness costs (i.e., altruistic punishment) can evolve is debated. Addressing this debate is beyond the scope of the current review and we refer interested readers elsewhere (27–29).

offenders, some of which require punishers to pay significant short-term costs, while others are less costly.

Based on the costs as well as the benefits of different tactics, we distinguish between two broad categories of punishment: *direct* punishment, which involves physical and verbal confrontation[2], and *indirect* (reputation-based) punishment, which involves gossip and ostracism. Punishing via direct confrontation has high costs—in terms of energetic expenditure, an increased risk of retaliation, and negative reputational consequences—but may also produce substantial benefits. Directly confronting offenders is more effective at changing their (future) behavior in ways that fit punishers' interests (32,33). In the context of status competition, direct confrontation may also bring some reputational benefits when there is value in building and maintaining a reputation of being a tough bargainer (34–36). In contrast, indirect punishment tactics have lower costs (37,38), because they allow punishers to remain anonymous (at least to offenders) and minimize the risk of retaliation. However, using indirect tactics of punishment is less effective at changing offenders' behavior, partly because offenders are unable to identify which of their behaviors has evoked punishment. Nevertheless, gossip and ostracism can impose significant reputational and relational costs on offenders (37,39). That is, offenders who are gossiped about tend to acquire a negative reputation, and thus are less likely to attract potential coalitional partners in future social interactions. Similarly, offenders who are ostracized suffer costs in terms of losing potentially valuable interaction opportunities.

Although field studies in small- and large-scale societies point to the key role that indirect, reputation-based tactics like gossip and ostracism play in promoting cooperation or

[2] Because we focus on peer-imposed punishment, we do not consider in detail other direct punishment tactics, such as fine imposition and imprisonment. Although these punishments clearly impose direct costs on offenders, in terms of reducing their wealth or compromising their freedom, they are typically decided upon and implemented by formal authorities and their representatives.

enforcing norms (40–43), the experimental literature has overwhelmingly focused on direct punishment via economic sanctioning. This focus can limit the ecological validity of research findings because (a) it remains unclear which real-world behaviors are captured by standard operationalizations of punishment in laboratory experiments, and (b) many frequent, consequential, but low-cost forms of cost imposition on offenders are often neglected. Here, we propose a framework that integrates a larger breadth of punishment and reputation-based tactics used to intervene against offenses. We suggest that the typology of intervention tactics we use here has the benefit of bridging strands of research on direct confrontation, gossip, and ostracism—behavioral phenomena which have often been studied separately. Considering the multiplicity of tactics that humans have available when deciding how to punish, along with the functions they serve (**section 2**) and the mechanisms that motivate them (**section 3**), highlights directions for future research on intervention against offenses, in the context of partner recalibration and partner choice, within interdependent relationships and social networks, and in daily life settings (**section 4**).

2. Common and Unique Social Functions of Distinct Punishment Tactics

Theoretical accounts of direct reciprocity, indirect reciprocity, and reputation-based partner choice suggest that multiple mechanisms can effectively promote and help sustain human cooperation (27,44,45). Mapping onto these accounts, empirical work has shown that people use a variety of tactics—direct confrontation, gossip, and ostracism—in response to non-cooperation and norm violations in real-world settings (40,43,46). For example, a study of responses to norm violations in a laboratory setting (47) found that around one quarter of witnesses directly intervened against a confederate who engaged in theft (cf. 48,49). A field experiment by Balafoutas and colleagues (50) found that a similar proportion of third-party observers directly

punished littering in a public space, though this rate of direct punishment dropped substantially when observers could indirectly punish the transgressor by withholding help. Consistently, a recent longitudinal study in daily life (51) showed that people intervene against offenses via various tactics, with gossip being the most frequent response, followed by direct confrontation, and social avoidance[3]. Together, these findings highlight the importance of studying the use of indirect reputation-based tactics alongside direct punishment tactics, to better understand how people intervene against offenses in daily settings and identify which goals punishment achieves.

Tactics of direct and indirect punishment are posited to serve similar broad functions: (a) promoting cooperation, (b) competing for resources and/or status, and (c) reducing inequality (see Figure 1). Seminal experiments have shown that punishment can be used to promote cooperation (6–8), though often at the expense of efficiency (for a review, see 25; cf. 53). More recent experiments have instead demonstrated that punishment is, in many cases, motivated by revenge (30,31; cf. 54), status concerns (35), or aversion to (disadvantageous) inequality (30,55). Traditional views of gossip and ostracism have emphasized the dark side of these tactics, seeing them as means to indirectly aggress against peers (37,39), and to impose status costs via reputation manipulation (56–58; especially in the context of resource competition, see 59 in this issue) . Recently, though, researchers have proposed broader conceptualizations of gossip that highlight its potential to strengthen social bonds and promote cooperative behaviors (38,60–62). In a similar vein, and despite research traditionally focusing on the negative emotional and social consequences of ostracism (63), experiments show that opportunities to choose some partners and exclude others can effectively promote cooperation (17,23). Although confrontation, gossip,

[3] It is worth noting that in this study the rate of direct punishment is much lower in situations that more closely resemble experimental tasks typically used to study punishment in the laboratory (i.e., second- and third-party punishment games; see also 52).

and ostracism can be used to achieve similar goals, each of these tactics has unique benefits and costs and may be additionally tailored to serve unique functions, which we articulate below.

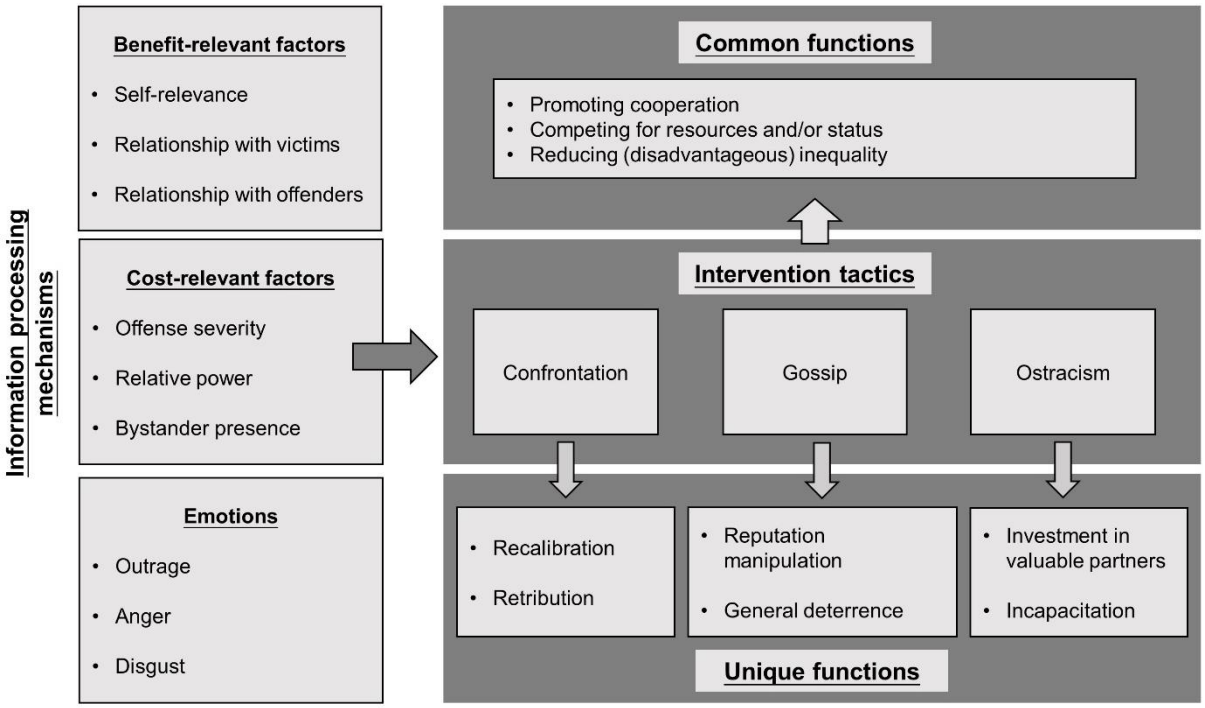

Figure 1. Information-processing mechanisms underlying decisions to intervene against offenses weigh recalibration benefits against retaliation costs. Direct punishment and indirect reputation-based tactics to intervene against offenses serve common and unique social functions.

2.1 The Unique Benefits and Costs of Direct Confrontation

Among the repertoire of available responses to offenses, direct confrontation seems better-tailored to recalibrate offenders' current and future behavior, in ways that benefit the punisher (32,33). This is because physically or verbally confronting offenders is the most immediate and effective way to stop ongoing transgressions and it allows the punished individuals to draw explicit links between instances of inappropriate behavior and the elicited punishment. This is not the case when offenses are met with gossip or ostracism, because these tactics do not convey information directly to the punished individuals about what they did

wrong. In contrast, verbal confrontation not only imposes costs on the offenders, but can communicate valuable information to them (64,65). For example, verbally confronting offenders can indicate which behaviors are perceived as offensive, how victimized parties are affected, and how the offenders should change their behavior to signal that they care about the punishers (66–68)[4]. Finally, compared to other intervention tactics, confrontation might be better suited to achieve retributive goals. Retribution involves a desire to balance or repay harm in a way that is proportionate to offense severity (71). Arguably, when using direct confrontation against offenders, punishers have more control over the immediate outcomes of their behavior, and they can adjust their responses more easily to fit the severity of the offense (72). In contrast, the outcomes of gossiping about an offender are often delayed and more uncertain, and the spread of information shared via gossip (to other individuals and even to the target of gossip) may be harder to control.

In sum, direct punishment via physical and verbal confrontation seems particularly well-tailored to achieve recalibration and retribution goals (see also 73), as compared to indirect reputation-based tactics to intervene against offenses. Notably, although directly confronting offenders can benefit punishers both in the short term, by putting an end to ongoing transgressions, and in the long term, by recalibrating offenders' future behavior to fit the punishers' interests, it comes with substantial costs. Direct confrontation requires time, effort, and energy; it bears the risk of counter-punishment and feuds (74,75); and it can result in incurring reputational costs (we return to this latter point in section 4).

[4] Relatedly, other work suggests that direct punishment (via economic sanctioning) is more effective at promoting cooperation when combined with communication (69,70).

2.2 The Unique Benefits and Costs of Gossip and Ostracism

Compared to direct confrontation, reputation-based tactics of gossip and ostracism seem well-suited to minimize the risk of retaliation when intervening against offenses, by obscuring the punishers' identity. However, as mentioned earlier, gossip and ostracism seem less effective at changing offenders' behavior, compared to direct confrontation. If gossip and ostracism are not primarily aimed at recalibration, what can they accomplish?

First, gossip plays a key role in facilitating cheater detection and partner choice. Indeed, people share and use reputational information to selectively cooperate with partners who have positive reputations (76–80) and avoid partners who have negative reputations (17,80,81). Further, although gossip itself may be less effective at changing targets' behaviors, the mere threat of being gossiped about by others motivates people to strive to build and maintain good reputations (16,18–20). Second, gossip is ideal for communicating about norms of acceptable behavior. It allows people to probe and safely test the limits of conventions, norms, and prescriptions, and facilitates the formation of strategic coalitions around moral values that fit personal interests (see 82,83) and/or group interests (84). Compared to other intervention tactics, gossip thus appears better suited to achieve general deterrence goals and it can be used as a sanction against norm violations (for a discussion of distinctions between punishment and sanctions, see 54,73). General deterrence involves a desire to prevent future offenses, not only from the same perpetrator, but also from third parties (i.e., *any* potential perpetrators) (71,85). Gossip can help achieve general deterrence by allowing people to coordinate with third parties and recruit punishment from them (81,86), lowering the otherwise high costs of uncoordinated direct punishment (see 87–89). Finally, gossip may represent one way for individuals to take

revenge by imposing strong *symbolic* costs on an offender, while reinstating their own image in the eyes of their community (54,73).

Compared to gossip, ostracism can be more costly, especially if used against valuable relationship partners, because it can result in missed interaction opportunities and severed social ties (51,67). Nevertheless, when used against partners with a lower relationship value, ostracism can achieve multiple goals. Given that there are cognitive and time constraints on the number and closeness of one's social relationships (90,91), the avoidance of offenders allows people to direct attention to more valuable social relationships. Indeed, reviews of the ethnographic record suggest that ostracism or avoidance is a common tactic to deal with norm violations (40,41,43), which limits the risk of conflict escalation. Finally, ostracism may be the most cost-effective way to incapacitate repeat offenders.

3. Information-Processing Mechanisms Underlying Distinct Punishment Tactics

Considering the differential costs and benefits associated with direct confrontation, gossip, and ostracism can be used to develop hypotheses about the putative information-processing mechanisms underlying direct punishment and indirect reputation-based tactics. Upon experiencing a norm violation, individuals first need to make decisions about whether to intervene or not and assess which tactics are available to them. If they decide to intervene, one possibility is that they then use whichever tactics they have available in an unconditional manner. An alternative possibility is that, upon deciding to intervene, people condition their choice of specific punishment tactics based on various situational factors. If so, what are the decision rules that they use to determine how to intervene against offenses, when both direct and indirect intervention tactics are available? Following previous research on punishment in daily life (51), we propose that individuals should consider and integrate (at least) two types of

information when deciding on which intervention tactic to use: (a) information about the benefits of recalibrating offenders' behavior and (b) information about the costs of being targeted by retaliation. We expect that, when the benefits of changing offenders' behavior are high, people will upregulate their use of direct punishment tactics; by contrast, when the costs of potential retaliation are high, people will upregulate their use of indirect, reputation-based tactics.

3.1 Factors that Shift Recalibration Benefits

We first consider several factors that can shift the benefits of punishment in terms of changing offenders' (future) behavior. One key factor that determines the benefits of intervening against an offense is the extent to which it has been personally harmful (i.e., the self-relevance of the offense). All else being equal, individuals have more to gain from deterring current and future offenses that are harmful for themselves. Indeed, multiple vignette studies have experimentally manipulated the self-relevance of offenses and found that people respond differently to violations victimizing themselves compared to those victimizing third parties. Specifically, offenses that are personally harmful are met with more direct, confrontational punishment (or with equally strong direct and indirect responses), whereas offenses that victimize third parties are preferentially met with indirect, less costly punishment tactics (92,93; but see, 94,95). Experience sampling studies on punishment in daily life settings (51,96) have found similar patterns, suggesting that self-relevant offenses evoke stronger desires to directly punish offenders, with whichever means possible, whereas other-relevant offenses are preferentially met with indirect punishment, via gossip or ostracism.

Importantly, most research to date has compared how people punish offenses that harm themselves to how they punish offenses that harm strangers. However, in real-world ecologies, people interact and experience offenses within diverse social relationships with kin, friends,

allies, ingroup members, and outgroup members. Considering the relationship context in which offenses take place is key for improving the ecological validity of research on punishment (51,52,97,98), and specifically for drawing accurate conclusions about the prevalence and use of punishment in the field. Indeed, recent work has taken promising steps in this direction, showing that people condition their punishment tactics on their relatedness and emotional closeness to victims. Consistent with the idea that people have more to gain from deterring offenses that are harmful for interdependent others (e.g., their kin, friends, and allies), offenses against close relationship partners (i.e., family and friends) evoke similar responses as self-relevant offenses, eliciting more costly confrontation than offenses against strangers (92,99). Moreover, offenses harming kin are met with harsher punishment than offenses harming friends, pointing to the role of relatedness and special obligations towards family in determining punishment (92,100,101).

In a similar vein, people might condition their punishment tactics on their relationship with offenders, especially when violations are self-relevant (i.e., when they are the victims and act as second-party punishers)[5]. Specifically, individuals may prefer directly confronting offenders whom they value highly rather than gossiping about them or ostracizing them (51). This prediction is based on several reasons. First, there is more to gain from adjusting the behavior of highly valued partners with whom one shares future interdependence. In contrast, if there is no expectation of future interactions with offenders, little can be gained by investing time and effort to recalibrate their behavior. Additionally, there is less uncertainty about how close others will respond to punishment. Finally, gossiping about valued partners can backfire if they find out, while ostracizing them can damage otherwise important social ties (57,67). Importantly,

[5] The relationship value of offenders might affect *third-party* punishment differently than what we suggest here (see 52,102).

additional prescriptions to intervene against offenses perpetrated by one's family or allies may apply in societies with strong kinship ties and norms of corporate responsibility (98,103).

3.2 Factors that Shift Retaliation Costs

When making decisions about how to intervene against offenses, people should not only consider the potential benefits in terms of recalibration, but also weigh the costs of receiving retaliation from offenders and their allies. Arguably, such costs of intervention differ depending on the severity of offenses, with more severe offenses being associated with a higher risk of retaliation. This is because offenders who have engaged in more morally wrong or harmful offenses may be perceived as more willing and able to retaliate if punished. Nevertheless, previous work has found that costly punishment increases with the severity of offenses, with people imposing harsher punishment against transgressions that are perceived as more severe (3,71), or transgressions that deviate more from group norms of cooperation (6). However, studies documenting how people use a broader array of intervention tactics in the field reveal that severe offenses are more often punished indirectly via gossip, ostracism, or withdrawal of help (51,104).

Another, more direct cue for assessing the risk of retaliation is the victim's relative power compared to that of the offender. Power can take many forms, including (a) one's privileged access to resources and the provision of benefits and costs, (b) one's asymmetric control over their own and others' outcomes, (c) one's influence derived from prestige, and (d) one's formidability based on their strength or other physical attributes (105–108). Individuals who experience high power relative to offenders—whatever the basis of this power—may be more willing to engage in direct, confrontational punishment (109–112), because they can afford the risk of retaliation. In contrast, individuals who find themselves in an unfavorable power position

relative to the offenders are expected to be more cautious against potential retaliation. Consistent with these ideas, people who feel less powerful are more likely to respond to norm violations by gossiping or avoiding the offenders, rather than by directly confronting them (51). Gossiping about transgressors also allows individuals who are less powerful to recruit punishment from third parties (40,81,86), potentially reducing individual costs of punishment and the risk of retaliation from powerful others.

In addition to the factors discussed earlier, the presence of bystanders may also influence the costs of third-party intervention against offenses, and thus the likelihood of intervention. For example, contrary to the notion of diffusion of responsibility, a quasi-experiment conducted on a train suggests that the silence norm is more likely to be enforced when there are more passengers in a train car (49). This could be because punishers expect others to take their side if the situation escalates, such that there are lower retaliation costs particularly when there are more bystanders present. Such a situation represents a volunteer's dilemma in which a single individual can maintain the (second-order) public good of silence in the train by punishing the norm violator (see also 89,113). Future research needs to consider how the presence of others influences not only the probability that someone intervenes, but also the use of specific types of intervention tactics. It is plausible that in situations that resemble the volunteer's dilemma, individuals observing a norm violation first use gossip to coordinate and then rely on only one person to directly confront an offender.

3.3 Emotions as Proximate Motivators of Punishment

So far, we have focused on the cognitive processes—whether conscious or unconscious—underlying decisions about how to punish offenses. Importantly, though, punishment is often motivated by negative emotions, including anger, disgust, and contempt

(96,114–116). Recent work has emphasized that different emotions may serve unique social functions (67,95,117), with anger and disgust motivating distinct responses to offenses. While anger is associated with approach-oriented, aggressive behaviors (68,118,119), disgust has been seen as motivating social avoidance (95,117,120) and efforts to signal condemnation to third parties (93,121). Consistent with these ideas, multiple vignette studies have shown that anger in response to offenses is specifically associated with inclinations to punish offenders directly, via physical and verbal confrontation (92,93,122). In contrast, moral disgust in response to the same offenses is associated with inclinations to punish offenders indirectly, via gossip and ostracism. These findings are corroborated by studies on punishment in daily life settings, showing that anger predicts both direct and indirect punishment responses, whereas disgust is specifically associated with gossip and ostracism (51). One potential explanation for why disgust motivates gossip against offenders is that sharing information about norm violations can effectively recruit subsequent ostracism from the receivers against the targets of gossip (81,86).

4. Addressing Current Debates and Carving Future Directions

In the preceding sections, we have drawn distinctions between multiple direct and indirect tactics to intervene against offenses—physical and verbal confrontation, gossip, and ostracism. In what follows, we describe how these distinctions can help address ongoing debates regarding the prevalence and functions of punishment, as well as the reputational consequences of third-party intervention against offenses.

First, various empirical studies have casted doubt on the generalizability and ecological validity of laboratory findings regarding the prevalence and use of punishment. Experimental research has shown that punishment may only promote cooperation under certain favorable conditions, such as when its cost-to-fine ratio is low (123) or when retaliation is not possible

(74,75). Further, reviews of the ethnographic record (41,43), as well as recent survey studies (51,52), suggest that punishment—especially as commonly operationalized in laboratory experiments (i.e., the imposition of monetary costs in response to offenses perpetrated by strangers against oneself or other strangers)—is rarely observed in the field. Delineating between direct punishment and indirect reputation-based tactics can facilitate comparisons between the lab and the field and ensure that experimental findings can be generalized to equivalent real-world situations. Conversely, pinpointing which real-world intervention tactics are of empirical interest can inform decisions about how to operationalize punishment in the lab.

Second, delineating between direct and indirect tactics can contribute to our understanding of when and how punishment serves cooperative versus competitive goals (for a detailed discussion, see 25). Confrontational punishment is largely evoked by offenses that harm oneself or close others (51,92,93,96); it involves aggressive inclinations, such as anger and revenge motives (30,31,116,124); and it can lead to feuds (74,75). Thus, while confrontational tactics may be favored in the context of status competition, using them among peers is often discouraged and, in some cases, even proscribed to ensure harmony within communities (125). In real-world settings, individuals often seem to prefer indirect, reputation-based tactics to deal with free-riding and other norm violations (50,51). Gossip, in particular, may be preferred over confrontational punishment because it allows individuals to first communicate about norms of acceptable behavior, and then coordinate their behavior with others, thus lowering the costs of intervention and the risk of conflicts.

Third, distinguishing between direct and indirect tactics can help address debates about the reputational consequences of third-party intervention against offenses. That is because the reputational consequences of intervention seem to vary depending on the tactics that are used to

impose costs on offenders. Experimental studies show that direct punishment, especially when imposed by third-party observers, effectively signals trustworthiness (126,127). However, when other means of intervention are available (e.g., helping the victims of transgressions), direct punishment loses some of its reliability as a signal of cooperativeness (128–130). Further, other findings cast doubt on the idea that second- and third-party punishment signal trustworthiness, and show that generous but not punitive individuals tend to be trusted more (131). Especially when intervention takes the form of physical or verbal confrontation it may even backfire, because confrontational individuals appear aggressive and are seen as motivated by selfish concerns (121,132). More work is needed to understand how observers perceive and judge intervention via reputation-based tactics of gossip and ostracism[6]. This is an especially interesting avenue for future work because different societies may deem different ways of intervening against offenses as more or less appropriate. A recent cross-cultural study of meta-norms (i.e., social norms about how people should treat norm violations) has provided some initial evidence that the appropriateness of confrontation, gossip, and ostracism differs across societies (125). Undoubtedly, understanding the ecological and cultural origins of variation in such meta-norms is a fascinating puzzle to be addressed by future research on cooperation.

Before closing, we turn to three additional recommendations for future research based on the work that we have reviewed here. First, our analysis suggests that the same intervention tactic can be used to achieve multiple purposes. To illustrate, gossip can be used for general deterrence (i.e., to recalibrate third parties' behavior), and can also be used to facilitate partner choice. Similarly, ostracism can represent an effort to recalibrate someone's behavior (e.g., in the

[6] On the one hand, gossip and ostracism can be used for prosocial purposes and may therefore have positive reputational consequences. On the other hand, gossip and ostracism can have deleterious effects on their targets (e.g., in the contexts of school bullying and public shaming on social media) and may therefore be negatively perceived by observers.

case of the ‘silent treatment’), but it is often used merely to navigate away from offenders and toward more valuable relationship partners. Future research on gossip and ostracism would benefit from studying when and how these reputation-based tactics are used for partner recalibration versus partner choice. Second, as noted earlier, third-party intervention in natural settings occurs within a rich relational context where the offender, the victim, and the third-party observer have varying degrees of interdependence with each other. Different structures of interdependence may affect when and how third parties choose to intervene against offenses (97,108,133). For example, people may be more prone to intervene against offenses that harm someone with whom they are mutually dependent (97), whereas they may be less prone to intervene against offenses that harm someone they have conflicting interests with. Considering interdependence relations and the properties of the networks that people are embedded in (e.g., network centrality and relational mobility; (134,135) is key to understanding third-party intervention. Likewise, our understanding of when people intervene against offenses in ecologically valid situations can be ameliorated via the use of a variety of field methods. We believe that by revisiting the rich ethnographic record and by using novel, experience sampling techniques, future work can document a variety of intervention tactics in real-world settings and provide valuable insights into the factors that shape the use of confrontation, gossip, and ostracism across social and cultural contexts.

Acknowledgements

We thank Jorge Peña for valuable comments on an earlier version of this manuscript. Catherine Molho acknowledges IAST funding from the French National Research Agency (ANR) under grant ANR-17-EURE-0010 (Investissements d'Avenir program). Junhui Wu acknowledges funding from the National Natural Science Foundation of China (71901028).

References

1. Gächter S, Schulz JF. Intrinsic honesty and the prevalence of rule violations across societies. Nature. 2016;531(7595):496–9.

2. Gelfand MJ, Raver JL, Nishii L, Leslie LM, Lun J, Lim BC, et al. Differences Between Tight and Loose Cultures: A 33-Nation Study. Science. 2011;332(6033):1100–4.

3. Hofmann W, Wisneski DC, Brandt MJ, Skitka LJ. Morality in everyday life. Science. 2014;345(6202):1340–3.

4. Balliet D, Mulder LB, Van Lange PAM. Reward, punishment, and cooperation: A meta-analysis. Psychol Bull. 2011;137(4):594–615.

5. Boyd R, Richerson PJ. Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethol Sociobiol. 1992;13(3):171–95.

6. Fehr E, Fischbacher U. The nature of human altruism. Nature. 2003;425(6960):785–91.

7. Fehr E, Gachter S. Cooperation and Punishment in Public Goods Experiments. Am Econ Rev. 2000;90(4):980–94.

8. Yamagishi T. The Provision of a Sanctioning System as a Public Good. J Pers Soc Psychol. 1986;51(1):110–6.

9. Henrich J, McElreath R, Barr A, Ensminger J, Barrett C, Bolyanatz A, et al. Costly Punishment Across Human Societies. Science. 2006;312(5781):1767–70.

10. Eriksson K, Strimling P, Andersson PA, Aveyard M, Brauer M, Gritskov V, et al. Cultural Universals and Cultural Differences in Meta-Norms about Peer Punishment. Manag Organ Rev. 2017;13(4):851–70.

11. Alexander R. The Biology of Moral Systems. Aldine de Gruyter; 1987.

12. Nowak MA, Sigmund K. Evolution of indirect reciprocity. Nature. 2005;437(7063):1291–8.

13. Barclay P. Trustworthiness and competitive altruism can also solve the “tragedy of the commons.” Evol Hum Behav. 2004;25(4):209–20.

14. Barclay P, Willer R. Partner choice creates competitive altruism in humans. Proc R Soc B Biol Sci. 2007;274(1610):749–53.

15. Sylwester K, Roberts G. Cooperators benefit through reputation-based partner choice in economic games. Biol Lett. 2010;6(5):659–62.

16. Feinberg M, Willer R, Keltner D. The Virtues of Gossip: Reputational Information Sharing as Prosocial Behavior. J Pers Soc Psychol. 2012;102(5):1015–30.

17. Feinberg M, Willer R, Schultz M. Gossip and Ostracism Promote Cooperation in Groups. Psychol Sci. 2014;25(3):656–64.

18. Piazza J, Bering JM. Concerns about reputation via gossip promote generous allocations in an economic game. Evol Hum Behav. 2008;29(3):172–8.

19. Wu J, Balliet D, Van Lange PAM. Reputation management: Why and how gossip enhances generosity. Evol Hum Behav. 2016;37(3):193–201.

20. Wu J, Balliet D, Van Lange PAM. When Does Gossip Promote Generosity? Indirect Reciprocity Under the Shadow of the Future. Soc Psychol Personal Sci. 2015;6(8):923–30.

21. Wu J, Balliet D, Lange PAMV. Gossip Versus Punishment: The Efficiency of Reputation to Promote and Maintain Cooperation. Sci Rep. 2016;6(1):1–8.

22. Grimalda G, Pondorfer A, Tracer DP. Social image concerns promote cooperation more than altruistic punishment. Nat Commun. 2016;7(1):12288.

23. Barclay P, Raihani N. Partner choice versus punishment in human Prisoner’s Dilemmas. Evol Hum Behav. 2016;37(4):263–71.

24. Rockenbach B, Milinski M. The efficient interaction of indirect reciprocity and costly punishment. Nature. 2006;444(7120):718–23.

25. Raihani NJ, Bshary R. Punishment: one tool, many uses. Evol Hum Sci. 2019;1:e12.

26. Clutton-Brock TH, Parker GA. Punishment in animal societies. Nature. 1995;373(6511):209–16.

27. West SA, El Mouden C, Gardner A. Sixteen common misconceptions about the evolution of cooperation in humans. Evol Hum Behav. 2011;32(4):231–62.

28. Boyd R, Gintis H, Bowles S, Richerson PJ. The evolution of altruistic punishment. Proc Natl Acad Sci. 2003;100(6):3531–5.

29. Lehmann L, Rousset F, Roze D, Keller L. Strong Reciprocity or Strong Ferocity? A Population Genetic View of the Evolution of Altruistic Punishment. Am Nat. 2007;170(1):21–36.

30. Bone JE, Raihani NJ. Human punishment is motivated by both a desire for revenge and a desire for equality. Evol Hum Behav. 2015;36(4):323–30.

31. Deutchman P, Bračič M, Raihani N, McAuliffe K. Punishment is strongly motivated by revenge and weakly motivated by inequity aversion. Evol Hum Behav. 2021;42(1):12–20.

32. Krasnow MM, Cosmides L, Pedersen EJ, Tooby J. What Are Punishment and Reputation for? PLOS ONE. 2012;7(9):e45662.

33. Krasnow MM, Delton AW, Cosmides L, Tooby J. Looking Under the Hood of Third-Party Punishment Reveals Design for Personal Benefit. Psychol Sci. 2016;27(3):405–18.

34. van Kleef GA, De Dreu CKW, Manstead ASR. The Interpersonal Effects of Anger and Happiness in Negotiations. J Pers Soc Psychol. 2004;86(1):57–76.

35. Yamagishi T, Horita Y, Mifune N, Hashimoto H, Li Y, Shinada M, et al. Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. Proc Natl Acad Sci. 2012;109(50):20364–8.

36. Raihani NJ, Bshary R. The reputation of punishers. Trends Ecol Evol. 2015 Feb 1;30(2):98–103.

37. Archer J, Coyne SM. An Integrated Review of Indirect, Relational, and Social Aggression. Personal Soc Psychol Rev. 2005;9(3):212–30.

38. Feinberg M, Cheng JT, Willer R. Gossip as an effective and low-cost form of punishment. Behav Brain Sci. 2012;35(1):25–25.

39. Campbell A. Staying alive: Evolution, culture, and women's intrasexual aggression. Behav Brain Sci. 1999;22(2):203–14.

40. Boehm C. Egalitarian Behavior and Reverse Dominance Hierarchy. Curr Anthropol. 1993;34(3):227–54.

41. Baumard N. Has punishment played a role in the evolution of cooperation? A critical review. Mind Soc. 2010;9(2):171–92.

42. Wiessner P. Norm enforcement among the Ju/'hoansi Bushmen : A case of strong reciprocity? Hum Nat. 2005;16(2):115–45.

43. Guala F. Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. Behav Brain Sci. 2012;35(1):1–15.

44. Nowak MA. Five Rules for the Evolution of Cooperation. Science. 2006;314(5805):1560–3.

45. Rand DG, Nowak MA. Human cooperation. Trends Cogn Sci. 2013;17(8):413–25.

46. Ostrom E. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge University Press; 1990. 308 p.

47. Baumert A, Halmburger A, Schmitt M. Interventions Against Norm Violations: Dispositional Determinants of Self-Reported and Real Moral Courage. Pers Soc Psychol Bull. 2013;39(8):1053–68.

48. Balafoutas L, Nikiforakis N. Norm enforcement in the city: A natural field experiment. Eur Econ Rev. 2012 Nov 1;56(8):1773–85.

49. Przepiorka W, Berger J. The Sanctioning Dilemma: A Quasi-Experiment on Social Norm Enforcement in the Train. Eur Sociol Rev. 2016 Jun 1;32(3):439–51.

50. Balafoutas L, Nikiforakis N, Rockenbach B. Direct and indirect punishment among strangers in the field. Proc Natl Acad Sci. 2014;111(45):15924–7.

51. Molho C, Tybur JM, Van Lange PAM, Balliet D. Direct and indirect punishment of norm violations in daily life. Nat Commun. 2020;11(1):3432.

52. Pedersen EJ, McAuliffe WHB, Shah Y, Tanaka H, Ohtsubo Y, McCullough ME. When and Why Do Third Parties Punish Outside of the Lab? A Cross-Cultural Recall Study. Soc Psychol Personal Sci. 2019;1948550619884565.

53. Gächter S, Renner E, Sefton M. The Long-Run Benefits of Punishment. Science. 2008;322(5907):1510–1510.

54. Giardini, F., & Conte, R. (2015). Revenge and conflict: Social and cognitive aspects. In Conflict and multimodal communication (pp. 71-89). Springer, Cham.

55. Raihani NJ, McAuliffe K. Human punishment is motivated by inequity aversion, not a desire for reciprocity. Biol Lett. 2012;8(5):802–4.

56. Beersma B, Van Kleef GA. Why People Gossip: An Empirical Analysis of Social Motives, Antecedents, and Consequences: Why People Gossip. J Appl Soc Psychol. 2012;42(11):2640–70.

57. Foster EK. Research on Gossip: Taxonomy, Methods, and Future Directions. Rev Gen Psychol. 2004;8(2):78–99.

58. Peters K, Fonseca MA. Truth, Lies, and Gossip. Psychol Sci. 2020;31(6):702–14.

59. Hess N, Hagen EH. Competitive gossip: The impact of domain, resource value, resource scarcity, and coalitions. Phil Trans R Soc B. (this issue).

60. Dores Cruz TD, Nieper AS, Testori M, Martinescu E, Beersma B. An Integrative Definition and Framework to Study Gossip. Group Organ Manag. 2021;46(2):252–85.

61. Dunbar RIM. Gossip in Evolutionary Perspective. Rev Gen Psychol. 2004;8(2):100–10.

62. Wu J, Balliet D, Van Lange PAM. Reputation, Gossip, and Human Cooperation. Soc Personal Psychol Compass. 2016;10(6):350–64.

63. Williams KD. Ostracism. Annu Rev Psychol. 2007;58(1):425–52.

64. Cushman F, Sarin A, Ho M. Punishment as communication. 2019. Available from: https://psyarxiv.com/wf3tz

65. Sarin A, Ho MK, Martin JW, Cushman FA. Punishment is Organized around Principles of Communicative Inference. Cognition. 2021;208:104544.

66. Delton AW, Robertson TE. How the mind makes welfare tradeoffs: evolution, computation, and emotion. Curr Opin Psychol. 2016;7:12–6.

67. Fischer AH, Roseman IJ. Beat them or ban them: The characteristics and social functions of anger and contempt. J Pers Soc Psychol. 2007;93(1):103–15.

68. Sell A, Tooby J, Cosmides L. Formidability and the logic of human anger. Proc Natl Acad Sci. 2009;106(35):15073–8.

69. Andrighetto G, Brandts J, Conte R, Sabater-Mir J, Solaz H, Villatoro D. Punish and Voice: Punishment Enhances Cooperation when Combined with Norm-Signalling. PLOS ONE. 2013;8(6):e64941.

70. Janssen MA, Holahan R, Lee A, Ostrom E. Lab Experiments for the Study of Social-Ecological Systems. Science. 2010;328(5978):613–7.

71. Carlsmith KM, Darley JM, Robinson PH. Why do we punish? Deterrence and just deserts as motives for punishment. J Pers Soc Psychol. 2002;83(2):284–99.

72. Molho C, Twardawski M, Fan L. What motivates direct and indirect punishment? Extending the “intuitive retributivism” hypothesis. Z Für Psychol. in press;

73. Giardini F, Andrighetto G, Conte R. A cognitive model of punishment. Proceedings of the Annual Meeting of the Cognitive Science Society. 2010.

74. Nikiforakis N. Punishment and counter-punishment in public good games: Can we really govern ourselves? J Public Econ. 2008;92(1):91–112.

75. Nikiforakis N, Engelmann D. Altruistic punishment and the threat of feuds. J Econ Behav Organ. 2011;78(3):319–32.

76. Abrahao B, Parigi P, Gupta A, Cook KS. Reputation offsets trust judgments based on social biases among Airbnb users. Proc Natl Acad Sci. 2017;114(37):9848–53.

77. Khadjavi M. Indirect Reciprocity and Charitable Giving— Evidence from a Field Experiment. Manag Sci. 2016;63(11):3708–17.

78. Sommerfeld RD, Krambeck H-J, Semmann D, Milinski M. Gossip as an alternative for direct observation in games of indirect reciprocity. Proc Natl Acad Sci. 2007;104(44):17435–40.

79. Wedekind C, Milinski M. Cooperation Through Image Scoring in Humans. Science. 2000;288(5467):850–2.

80. Diekmann A, Jann B, Przepiorka W, Wehrli S. Reputation Formation and the Evolution of Cooperation in Anonymous Online Markets. Am Sociol Rev. 2014;79(1):65–85.

81. Dores Cruz TD, Thielmann I, Columbus S, Molho C, Wu J, Righetti F, et al. Gossip and Reputation in Everyday Life. Phil Trans R Soc B. (this issue).

82. DeScioli P, Kurzban R. Mysteries of morality. Cognition. 2009 Aug 1;112(2):281–99.

83. DeScioli P, Kurzban R. A solution to the mysteries of morality. Psychol Bull. 2013;139(2):477–96.

84. Beersma, B., Van Kleef, G.A. How the grapevine keeps you in line: Gossip increases contributions to the group. Soc Psychol Personal Sci. 2011;2:642–9.

85. Twardawski M, Tang KTY, Hilbig BE. Is It All About Retribution? The Flexibility of Punishment Goals. Soc Justice Res. 2020;195–218.

86. Dores Cruz TD, van der Lee R, Beersma B. Gossip about Coronavirus: Infection status and norm adherence shape social responses. Group Process Intergroup Relat. 2021 Jun;24(4):658–79.

87. Boyd R, Gintis H, Bowles S. Coordinated Punishment of Defectors Sustains Cooperation and Can Proliferate When Rare. Science. 2010;328(5978):617–20.

88. Molleman L, Kölle F, Starmer C, Gächter S. People prefer coordinated punishment in cooperative interactions. Nat Hum Behav. 2019;3(11):1145–53.

89. Przepiorka W, Diekmann A. Individual heterogeneity and costly punishment: a volunteer's dilemma. Proc R Soc B Biol Sci. 2013;280(1759):20130247.

90. Stiller J, Dunbar RIM. Perspective-taking and memory capacity predict social network size. Soc Netw. 2007;29(1):93–104.

91. Sutcliffe A, Dunbar R, Binder J, Arrow H. Relationships and the social brain: Integrating psychological and evolutionary perspectives. Br J Psychol. 2012;103(2):149–68.

92. Lopez LD, Moorman K, Schneider S, Baker MN, Holbrook C. Morality is relative: Anger, disgust, and aggression as contingent responses to sibling versus acquaintance harm. Emotion. 2019;

93. Molho C, Tybur JM, Güler E, Balliet D, Hofmann W. Disgust and Anger Relate to Different Aggressive Responses to Moral Violations. Psychol Sci. 2017;28(5):609–19.

94. FeldmanHall O, Sokol-Hessner P, Van Bavel JJ, Phelps EA. Fairness violations elicit greater punishment on behalf of another than for oneself. Nat Commun. 2014;5(1):5306.

95. Hutcherson CA, Gross JJ. The moral emotions: A social–functionalist account of anger, disgust, and contempt. J Pers Soc Psychol. 2011;100(4):719–37.

96. Hofmann W, Brandt MJ, Wisneski DC, Rockenbach B, Skitka LJ. Moral Punishment in Everyday Life. Pers Soc Psychol Bull. 2018;44(12):1697–711.

97. Horne C. The Rewards of Punishment: A Relational Theory of Norm Enforcement. Stanford University Press; 2009.

98. Moya C, Fessler D, Henrich J, Zhao W, Barrett HC, et al. Norm enforcement in small-scale societies depends on coordinated third party responses and pre-existing relationships. Unpublished Manuscript.

99. Pedersen EJ, McAuliffe WHB, McCullough ME. The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically. J Exp Psychol Gen. 2018;147(4):514–44.

100. McManus RM, Kleiman-Weiner M, Young L. What We Owe to Family: The Impact of Special Obligations on Moral Judgment. Psychol Sci. 2020;31(3):227–42.

101. Lieberman D, Linke L. The Effect of Social Category on Third Party Punishment. Evol Psychol. 2007 Apr 1;5(2):147470490700500.

102. Delton AW, Krasnow MM. The psychology of deterrence explains why group membership matters for third-party punishment. Evol Hum Behav. 2017;38(6):734–43.

103. Henrich J. The WEIRDest People in the World: How the West Became Psychologically Peculiar and Particularly Prosperous. Farrar, Straus and Giroux; 2020.

104. Balafoutas L, Nikiforakis N, Rockenbach B. Altruistic punishment does not increase with the severity of norm violations in the field. Nat Commun. 2016;7(1):13327.

105. Cheng JT. Dominance, prestige, and the role of leveling in human social hierarchy and equality. Curr Opin Psychol. 2020;33:238–44.

106. Garfield ZH, Hubbard RL, Hagen EH. Evolutionary Models of Leadership. Hum Nat. 2019;30(1):23–58.

107. Durkee PK, Lukaszewski AW, Buss DM. Psychological foundations of human status allocation. Proc Natl Acad Sci. 2020 Sep 1;117(35):21235–41.

108. Kelley HH, Holmes JG, Kerr NL, Reis HT, Rusbult CE, Lange PAMV. An Atlas of Interpersonal Situations. Cambridge University Press; 2003.

109. Gordon DS, Lea SEG. Who Punishes? The Status of the Punishers Affects the Perceived Success of, and Indirect Benefits From, "Moralistic" Punishment. Evol Psychol. 2016;14(3):1474704916658042.

110. Chierchia G, Lesemann FHP, Snower D, Vogel M, Singer T. Caring Cooperators and Powerful Punishers: Differential Effects of Induced Care and Power Motivation on Different Types of Economic Decision Making. Sci Rep. 2017;7(1):11068.

111. Redhead D, Dhaliwal N, Cheng JT. Taking charge and stepping in: Individuals who punish are rewarded with prestige and dominance. Soc Personal Psychol Compass. 2021;15(2):e12581.

112. Molho C, Balliet D, Wu J. Hierarchy, Power, and Strategies to Promote Cooperation in Social Dilemmas. Games. 2019;10(1):12.

113. Raihani NJ, Bshary R. The evolution of punishment in n-player public goods games: A volunteer's dilemma. Evolution. 2011;65(10):2725–8.

114. Chapman HA, Kim DA, Susskind JM, Anderson AK. In Bad Taste: Evidence for the Oral Origins of Moral Disgust. Science. 2009;323(5918):1222–6.

115. Hopfensitz A, Reuben E. The Importance of Emotions for the Effectiveness of Social Punishment. Econ J. 2009;119(540):1534–59.

116. Seip EC, Van Dijk WW, Rotteveel M. Anger motivates costly punishment of unfair behavior. Motiv Emot. 2014;38(4):578–88.

117. Tybur JM, Lieberman D, Kurzban R, DeScioli P. Disgust: Evolved function and structure. Psychol Rev. 2013;120(1):65–84.

118. Carver CS, Harmon-Jones E. Anger is an approach-related affect: Evidence and implications. Psychol Bull. 2009;135(2):183–204.

119. Harmon-Jones E, Allen JJB. Anger and frontal brain activity: EEG asymmetry consistent with approach motivation despite negative affective valence. J Pers Soc Psychol. 1998;74(5):1310–6.

120. Curtis V, Biran A. Dirt, Disgust, and Disease: Is Hygiene in Our Genes? Perspect Biol Med. 2001;44(1):17–31.

121. Kupfer TR, Giner-Sorolla R. Communicating Moral Motives: The Social Signaling Function of Disgust. Soc Psychol Personal Sci. 2017;8(6):632–40.

122. Tybur JM, Molho C, Cakmak B, Cruz TDDD, Singh GD, Zwicker M. Disgust, anger, and aggression: Further tests of the equivalence of moral emotions. Collabra Psychol. 2020;6(1):34.

123. Egas M, Riedl A. The economics of altruistic punishment and the maintenance of cooperation. Proc R Soc B Biol Sci. 2008;275(1637):871–8.

124. Mischkowski D, Glöckner A, Lewisch P. From spontaneous cooperation to spontaneous punishment – Distinguishing the underlying motives driving spontaneous behavior in first and second order public good games. Organ Behav Hum Decis Process. 2018;149:59–72.

125. Eriksson K, Strimling P, Gelfand M, Wu J, Abernathy J, et al. Perceptions of the appropriate response to norm violation in 57 societies. Nat Commun. 2021;12(1):1–11.

126. Barclay P. Reputational benefits for altruistic punishment. Evol Hum Behav. 2006;27(5):325–44.

127. Jordan JJ, Hoffman M, Bloom P, Rand DG. Third-party punishment as a costly signal of trustworthiness. Nature. 2016;530(7591):473–6.

128. Raihani NJ, Bshary R. Third-party punishers are rewarded, but third-party helpers even more so. Evolution. 2015;69(4):993–1003.

129. Jordan JJ, Rand DG. Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. J Pers Soc Psychol. 2020;118(1):57–88.

130. Dhaliwal NA, Patil I, Cushman F. Reputational and cooperative benefits of third-party compensation. Organ Behav Hum Decis Process. 2021;164:27–51.

131. Przepiorka W, Liebe U. Generosity is a sign of trustworthiness—the punishment of selfishness is not. Evol Hum Behav. 2016 Jul 1;37(4):255–62.

132. Eriksson K, Andersson PA, Strimling P. Moderators of the disapproval of peer punishment. Group Process Intergroup Relat. 2016;19(2):152–68.

133. Balliet DP, Tybur JM, Lange PAM van. Functional Interdependence Theory: An evolutionary account of social situations. Personal Soc Psychol Rev. 2017;21(4):361–88.

134. Breza E, Chandrasekhar A, Larreguy H. Network centrality and informal institutions: Evidence from a lab experiment in the field. Working Paper No 562 Stanford University. 2016.

135. Roos P, Gelfand M, Nau D, Carr R. High strength-of-ties and low mobility enable the evolution of third-party punishment. Proc R Soc B Biol Sci. 2014;281(1776):20132661.