Evolution and Human Behavior 44 (2023) 50–59

Contents lists available at ScienceDirect

Evolution and Human Behavior

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

Why punish cheaters? Those who withdraw cooperation enjoy better reputations than punishers, but both are viewed as difficult to exploit

Sakura Arai [a,b,c,*], John Tooby [a,d], Leda Cosmides [a,b]

[a] Center for Evolutionary Psychology, University of California, Santa Barbara, CA 93106, United States
[b] Department of Psychological & Brain Sciences, University of California, Santa Barbara, CA 93106, United States
[c] Brain Science Institute, Tamagawa University, 6-1-1 Tamagawa-gakuen, Machida, Tokyo, 194-8610, Japan
[d] Department of Anthropology, University of California, Santa Barbara, CA 93106, United States

ARTICLE INFO

Keywords:
Negative sanction
Punishment
Conditional cooperation
Reputation
Partner choice

ABSTRACT

Negatively sanctioning cheaters promotes cooperation. But do all negative sanctions have the same consequences? In dyadic cooperation, there are two ways that cooperators can sanction failures to reciprocate: by inflicting punishment or withdrawing cooperation. Although punishment can be costly, it has been proposed that this cost can be recouped if punishers acquire better reputations than non-punishers and, therefore, are favored as cooperation partners. But the evidence so far is mixed, and nothing is known about the reputations of those who sanction by withdrawing cooperation. Here, we test two novel hypotheses about how inflicting negative sanctions affects the reputation of the sanctioner: (i) Those who withdraw cooperation are evaluated more favorably than punishers, and (ii) both sanctioners are viewed as less exploitable than non-sanctioners. Observers (US online convenience sample, $n = 246$) evaluated withdrawers as more cooperative and less vengeful than punishers and preferred withdrawers as a partner. Sanctioners were also viewed as more difficult to exploit than non-sanctioners, with no difference between punishers and withdrawers. The results were the same when punishment was costly (US college sample, $n = 203$) with one exception: Costly punishers, who lost their payoffs by punishing, were viewed as more exploitable than withdrawers. Our results indicate that withdrawing cooperation has advantages over punishing: Withdrawers are favored as cooperative partners while gaining a reputation as difficult to exploit. The reputational consequences of the three responses to defectors—punishing, withdrawing cooperation, and not sanctioning at all—were opposite to those predicted by group selection models.

1. Introduction

Negatively sanctioning cheaters promotes cooperation. But there are two ways of sanctioning partners who fail to reciprocate: by withdrawing cooperation or inflicting punishment. Punishment—inflicting a cost that reduces the payoff of a cheater—has been shown to successfully sustain cooperation (Fehr & Gächter, 2000, 2002; Yamagishi, 1986). But inflicting punishment is sometimes costly to the punisher as well (Clutton-Brock & Parker, 1995), leading theorists to ask how selection could have favored punishment as a means of sanctioning cheaters (Panchanathan & Boyd, 2004; Tooby, Cosmides, & Price, 2006).

Several researchers have proposed that the cost of inflicting punishment can be recouped if punishers acquire reputations as better cooperative partners than non-punishers, thereby attracting (or retaining) more rewarding partners for future interactions (Barclay, 2006; Horita, 2010; Kiyonari & Barclay, 2008; Ozono & Watabe, 2012; Raihani & Bshary, 2015a). Tests of this hypothesis have generated mixed results. Some studies found that punishers were seen as more trustworthy and received more benefits than non-punishers (Barclay, 2006; dos Santos, Rankin, & Wedekind, 2013; Jordan, Hoffman, Bloom, & Rand, 2016; Nelissen, 2008; Raihani & Bshary, 2015b), but others found that punishers were seen as less trustworthy and reaped no advantage over non-punishers (Balafoutas, Nikiforakis, & Rockenbach, 2014; Barclay & Raihani, 2016; Bone, Wallace, Bshary, & Raihani, 2016; Fehr & Rockenbach, 2003; Kiyonari & Barclay, 2008; Przepiorka & Liebe, 2016). Many variables could account for these conflicting results (Horita, 2010; Mifune, Li, & Okuda, 2020; Ozono & Watabe, 2012; Raihani & Bshary, 2015a). Context may matter, for example, because

* Corresponding author at: Brain Science Institute, Tamagawa University, 6-1-1 Tamagawa-gakuen, Machida, Tokyo, 194-8610, Japan.
E-mail address: sakura.arai.psych@gmail.com (S. Arai).

https://doi.org/10.1016/j.evolhumbehav.2022.10.002
Received 13 April 2022; Received in revised form 14 October 2022; Accepted 15 October 2022
Available online 22 October 2022

group cooperation poses different problems than dyadic social exchange.

When three or more individuals cooperate to achieve a common goal and share the resulting benefits, punishment is the only way to selectively sanction free riders: Withdrawing cooperation from a free rider also withdraws it from cooperators, who contributed generously to the group project (Tooby et al., 2006). And if punishing succeeds in reforming the free rider, everyone else in the group benefits from associating with the punisher—leading some researchers to call this "altruistic punishment" (Bowles & Gintis, 2004; Fehr, Fischbacher, & Gächter, 2002; but see Krasnow, Delton, Cosmides, & Tooby, 2015). If altruistic punishment creates positive externalities for other group members, the punisher may gain a reputation as a valuable partner to include in future group projects.

None of this is true for dyadic cooperation. When two individuals reciprocally exchange benefits, withdrawing cooperation does selectively sanction the defector: It does not harm other, more trustworthy people with whom you may cooperate in the future. Either sanction—withdrawing cooperation or inflicting punishment—creates incentives for the defector to start cooperating with the sanctioner. But the defector's reformed behavior need not create positive externalities for anyone outside the dyad: The sanctioned person may continue to defect when cooperating with other people.

Two-person cooperation—reciprocating favors, swapping the fruits of foraging, exchanging goods and services—is ubiquitous in humans, and far more frequent than group cooperation. These two contexts are sufficiently different that we focused on the reputational consequences of sanctioning in just one of them: dyadic cooperation, where reciprocation is possible.

We investigated the reputational consequences of three possible responses a cooperator could have to a partner's failure to reciprocate: inflicting punishment, withdrawing cooperation, and not sanctioning at all. Strategies that cooperate conditionally are evolutionarily stable against strategies that defect (Tooby et al., 2006; Trivers, 1971; Williams, 1966), but many negative sanctions can incentivize a defecting partner to cooperate. Punishment does so by reducing the immediate payoff the partner gains by defecting. An alternative sanctioning strategy is to withdraw the benefits of cooperation: One can refrain from delivering additional benefits until the partner resumes cooperation (as TIT FOR TAT does; Axelrod & Hamilton, 1981) or switch to more rewarding partners until the defector reforms (Hammerstein & Noë, 2016; Tooby et al., 2006).

Very few studies have directly compared behavior in response to these two negative sanctions: punishing versus withdrawing cooperation (for an exception, see Barclay & Raihani, 2016). Moreover, we can find no studies of the reputational consequences of withdrawing cooperation, even though this was the most widely studied method of sanctioning in the early literature on the evolution of cooperation (e.g., Axelrod, 1984). The reputation attributed to those who withdraw cooperation has not been compared to that of punishers—or to the reputation of those who do not sanction at all.

We examined how these two methods of sanctioning influence the inferences observers make about the sanctioner's character and traits—the various reputation*s* (plural) that observers attribute to the sanctioner. The colloquial use of *reputation* implies a unitary dimension: Your reputation can become better or worse. But people are routinely evaluated on many different traits: Alex may have a reputation for being generous, a reputation for being lazy, and a reputation for being vengeful, for example. These need not merge to form a single "reputation." And, even if they do, these separate reputations should remain stored in the observer's memory, because which is most relevant depends on the situation a decision-maker is facing (Klein, Cosmides, Tooby, & Chance, 2002). Indeed, research on social cognition shows that the mind spontaneously infers many different traits rapidly, even from thin information (Funder & Sneed, 1993; Klein et al., 2009), and stores summary representations of each (Klein et al., 2009).

Here we test two previously unexamined hypotheses about the inferences people draw from a cooperator's response to a partner who defects. The first hypothesis regards the reputations of cooperators who respond by imposing negative sanctions: withdrawers and punishers. In the two studies reported herein, *withdrawer* refers to a cooperator who sanctions by not providing benefits to the defector in the next round, and *punisher* refers to a cooperator who sanctions by removing resources from the defector in the next round. We propose that withdrawers will acquire reputations for being more cooperative than punishers—they will be seen as, e.g., more generous, trustworthy, and forgiving. As a result, observers will prefer withdrawers to punishers as potential partners (Barclay, 2013; Roberts et al., 2021).

Why? Both withdrawers and punishers signal a willingness to sanction a defection, but withdrawers do so without reducing the payoff to a potentially well-intentioned cooperator. Even reliable cooperative partners will sometimes fail to reciprocate due to mistakes or bad luck (Delton, Cosmides, Guemo, Robertson, & Tooby, 2012); deciding whether a failure reveals a disposition to cheat versus a mistake is a judgment made under uncertainty. Because they are robust to mistakes, strategies that require more evidence before sanctioning a partner, such as TIT FOR TWO TATS, outcompete strategies that sanction immediately in agent-based simulations (Axelrod, 1984). As a result, they maintain cooperation with a partner instead of triggering cycles of mutual defection. In Study 1, sanctions are immediate in both cases and neither cooperator donates resources to their partner in the round following defection. But punishers take back what they gave whereas withdrawers do not. The partner—who may have made a mistake—retains the payoff provided by the withdrawer in the first round. This should lead observers to see the withdrawer as more generous and less vengeful than the punisher.

The second hypothesis addresses the reputational cost of *not* imposing negative sanctions when a partner defects. In both studies, *non-sanctioners* are cooperators who respond to defection by continuing to provide benefits to their partner. We propose that non-sanctioners will acquire a reputation for being more exploitable than those who impose negative sanctions, whether the sanctioners are punishers or withdrawers.

Why? Motivations to sanction defections could have been favored by selection if their average effect was to either increase benefits to the sanctioner and/or prevent losses. Previous research on the reputational consequences of sanctioning has focused on whether punishers gain more benefits from cooperation than non-sanctioners do (Balafoutas et al., 2014; Barclay, 2006; Barclay & Raihani, 2016; Bone et al., 2016; dos Santos et al., 2013; Fehr & Rockenbach, 2003; Horita, 2010; Jordan et al., 2016; Kiyonari & Barclay, 2008; Mifune et al., 2020; Nelissen, 2008; Ozono & Watabe, 2012; Przepiorka & Liebe, 2016; Raihani & Bshary, 2015a, 2015b). But only a handful of studies have examined the possibility that sanctioning protects the sanctioner from further losses (Delton & Krasnow, 2017; Hilbe & Traulsen, 2012; Krasnow, Delton, Cosmides, & Tooby, 2016; Yamagishi et al., 2009). The few studies that do suggest that motivations to sanction were designed to deter further maltreatment by the defector or other observers. In this view, the cost of *not* sanctioning defections is gaining a reputation for being exploitable, which invites mistreatment. If selection for preventing losses designed motivations to sanction defectors, then observers will view non-sanctioners as more exploitable than sanctioners.

We tested these two hypotheses by having participants observe how a cooperator responded to a failure to reciprocate. After, they made inferences about the character and traits of withdrawers, punishers, and non-sanctioners.

- H1: Withdrawers will be evaluated more favorably as a cooperation partner than punishers.
- H2: Sanctioners—withdrawers and punishers—will be evaluated as less exploitable than non-sanctioners.

Inflicting punishment was cost-free in Study 1 and costly in Study 2.

2. Study 1

In most theoretical and empirical work on the reputational consequences of punishment, the punisher pays a cost to reduce the payoff of a defector. Punishing can indeed be costly: This can be deduced from observations of non-human organisms, where punishment may entail energetic costs or the risk of injury—e.g., engaging in a physical fight with the defector, or chasing the defector off over long distances (Clutton-Brock & Parker, 1995).[1]

Reducing a defector's payoff need not be costly, however; the cost of punishment is sometimes negligible. In many species, displays of relative formidability, which are low cost by design, establish resource holding potential in advance of conflicts (Hammerstein & Parker, 1982). When both parties know the cooperator is more formidable than the defector, the defector may cede the resource upon being approached by the more formidable cooperator. If not, a reminder display may suffice. The cost of punishment declines sharply when cooperators coordinate to jointly inflict it (Boyd, Gintis, & Bowles, 2010), and language facilitates coordinated punishment (Wrangham, 2019). Talk is cheap (literally): Language—and nonverbal communication—can be leveraged to reduce a defector's payoffs. Telling the defector you want your fair share can be effective, as are cost-free signals of disapproval in economic games (Masclet, Noussair, Tucker, & Villeval, 2003). When toddlers collaborate to gain resources, the child who got more creates parity when the other child just shows that he got less (Hamann, Warneken, Greenberg, & Tomasello, 2011; Tomasello, 2009; Warneken, Lohse, Melis, & Tomasello, 2011). Gossip can harm the defector's reputation: Telling others you were treated badly may lead them to devalue the defector (Sznycer et al., 2016).

Even when punishment is costly, the cost can be recovered if resources taken from the defector go to the cooperator. The punisher can recoup the investment lost by the defector's failure to reciprocate—or even take more, to impose an additional penalty for cheating.

In Study 1, there is no cost to sanctioning, but punishers reclaim what they lost and withdrawers do not. It addresses the reputational consequences of punishment that does not entail spite (incurring a cost to inflict a cost).

2.1. Methods

This study was pre-registered prior to data collection (https://osf.io/zwb26/?view_only=ca275e8bf4664b1381d086227bed0274)[2] and approved by the Institutional Review Board at University of California, Santa Barbara. Data is available through the Open Science Framework (https://osf.io/yg56s/?view_only=b25462dc0b64411285e28460b02dd973). See Supplementary Material S1 for materials.

2.1.1. Participants

Participants were English speakers in the United States (n = 246,[3] 48.78% female, $M_{age} = 29$, $SD_{age} = 9$)[4] recruited via Prolific. Those who wished to participate in the study first completed a written informed consent form. The online study lasted about 8 min, and participants received US $1.28 for their participation.

2.1.2. Presenting a cooperator's response to defection

Participants were instructed that they would observe two individuals repeatedly interact in a Dictator Game with Taking Option (DGwT) (List, 2007). It was explained that there are two roles: giver and receiver. Both individuals are given $5 at the beginning of a round; then the giver receives an additional endowment of $5. The giver decides either to share this endowment with the receiver (up to $5) or to take money from the receiver (up to $5), both in $1 increments. After the giver's decision, the two switch roles and interact again.

After the explanation, participants observed two individuals, Alex and Casey, play three rounds of DGwT. (These names were chosen because they can apply to any gender; in reporting results, we will refer to both as "she" for ease of exposition.) Participants were told that Alex and Casey knew that they would interact repeatedly (participants did not know for how many rounds). In round 1, where Alex was the giver and Casey was the receiver, Alex gave $5 to Casey. In round 2, where Casey became the giver, Casey gave $0 to Alex, the receiver. Notice that Alex cooperated in round 1, and Casey failed to reciprocate in round 2.

In round 3, Alex became the giver again. Participants observed Alex make one of three responses in round 3 (between-subjects conditions):

- Punish: Alex took $5 from Casey
- Withdraw cooperation: Alex gave $0 to Casey
- No negative sanction (keep cooperating): Alex gave $5 to Casey again.

Terms such as "cooperation" and "punishment" were not used in the instructions to participants. Participants who did not understand or remember Alex's response were excluded from the study (see Supplementary Material S1).

2.1.3. Evaluating reputations

After observing the interaction, participants evaluated Alex on 24 adjectives: exploitable, weak, gullible, unwise, incompetent, vengeful, aggressive, impulsive, cowardly, frightened, mean, careless, dependable, likable, forgiving, generous, considerate, cooperative, trustworthy, honorable, friendly, kind, fair, and emotionally-stable. The order of the adjectives was randomized. Adjectives were taken from previous research (Barclay, 2006; Delton et al., 2012; Kiyonari & Barclay, 2008; Nelissen, 2008) or unanimously nominated by the authors. Each adjective was rated on a 7-point Likert scale (from 1: "Not at all" to 7: "Extremely"). Participants also rated how much they would like to interact with Alex in a DGwT on a 5-point Likert scale (from 1: "Not at all" to 5: "Extremely").[5]

[1] Assuming punishment is costly may stem from the intuition that defectors can retaliate against a sanctioner in real life. But that can happen to a withdrawer as well as a punisher; trade wars between nations are an example.

[2] Studies 1 and 2 deviate from the pre-registration in two ways. (i) To avoid second-guessing which traits participants view as positive or negative, we did not assign adjectives or reputations a priori to a positive versus a negative cluster before factor analyzing. (ii) We simplified the hypotheses to reflect this. Results are qualitatively the same regardless of how the analysis is done.

[3] The sample size was pre-registered and determined, using G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007) to obtain .80 power to detect a medium-sized effect (Cohen's $f = .20$) in one-way ANOVA at the .05 alpha error probability.

[4] We did not ask participants their demographic information except for their gender identity and age. Adding these variables in the analyses did not change the results reported in this paper.

[5] We also asked two additional questions: (a) what participants would do if they had to interact with Alex and (b) what they would do to Casey if they were Alex. See Supplementary Material S1 and S4 for details.

2.2. Results

2.2.1. Summary reputations

Data were analyzed using R 4.0.3 (R Core Team, 2020). First, we created summary reputations by using factor analysis to group related adjective ratings. Three factors were obtained on 24 adjective ratings, using the factanal function in R (R Core Team, 2020) with promax rotation, explaining 53.3% of the total variance. The number of factors was corroborated by parallel analysis using the fa.parallel function in the R package psych (Revelle, 2021).

We obtained three summary reputations by averaging the adjective ratings for each factor. Eleven adjectives, such as cooperative, trustworthy, considerate, and generous, composed a summary reputation for being *cooperative* (Cronbach's $\alpha = .93$) (see Supplementary Table S1 for other adjectives and factor loadings). Four adjectives—vengeful, aggressive, mean, and (un)forgiving (reverse-coded forgiving)—composed a summary reputation for being *vengeful* ($\alpha = .83$). Nine adjectives, such as exploitable, gullible, weak, and unwise, composed a summary reputation for being *exploitable* ($\alpha = .87$). The summary reputation for being *cooperative* was negatively correlated with the other two: $r\ (244) = -.65$, $p = 10^{-16}$ with *vengeful*; $-.25$, $p = 10^{-5}$ with *exploitable*. The correlation between the *vengeful* and *exploitable* summaries was positive, but not significant (.10, $p = .125$).

2.2.2. Reputational outcomes

We compared the reputations of Alex as a punisher, withdrawer, and non-sanctioner by conducting one-way ANOVAs and post-hoc pairwise comparisons on three summary reputations, using the aov and TukeyHSD functions in R (R Core Team, 2020).

There were significant differences in how cooperative (F [2, 243] = 40.4, $p = 10^{-16}$) and vengeful (F [2, 243] = 139.7, $p = 10^{-16}$) participants found the three responders. Supporting H1, withdrawers were evaluated as more cooperative (5.25 vs. 4.60, $p = 10^{-5}$) and less vengeful (3.33 vs. 4.09, $p = 10^{-7}$) than punishers (Fig. 1a and b). Non-sanctioners were seen as more cooperative ($p = 10^{-5}$) and less vengeful ($p < 10^{-16}$) than withdrawers. People found punishers the least cooperative and the most vengeful.

There was a significant difference in how exploitable (F [2, 243] = 5.27, $p = .006$) participants found the three responders. People found non-sanctioners the most exploitable (Fig. 1c). Supporting H2, punishers were evaluated as less exploitable than non-sanctioners (2.86 vs. 3.27, $p = .028$); so were withdrawers (2.80, $p = .009$). However, participants found withdrawers and punishers equally difficult to exploit ($p = .91$).

2.2.3. Partner choice

Additionally, we analyzed the single-item rating of how much participants would like to interact with the three responders. There were significant differences in how desirable they were viewed as a potential cooperation partner (F [2, 243] = 22.55, $p = 10^{-9}$). Punishers were least preferred: they were rated lower than withdrawers (3.48 vs. 4.17, $p = 10^{-5}$) and non-sanctioners (4.43, $p = 10^{-9}$) (Fig. 1d). But preferences were the same for withdrawers and non-sanctioners ($p = .19$).

The partner choice preference was positively correlated with the summary reputation for being cooperative (r [244] = .74, $p = 10^{-16}$) and negatively with the ones for being vengeful ($-.49$, $p = 10^{-16}$) and exploitable ($-.24$, $p = .0002$).

When controlling for other reputations and which response Alex made, only the summary reputation for being cooperative ($\beta = .69$, $p < 10^{-16}$) significantly increased how much participants wanted to interact with Alex (multiple regression using the lm function in R [R Core Team, 2020]). (See Supplementary Table S2 for a full model.) (The same was true when reputations were the only predictors in the model; model fit [AIC] was slightly better when Alex's responses were also included as predictors).

2.3. Discussion

Alex's reputation for cooperativeness differed across conditions, even though she always gave generously to Casey in the first round. She was seen as least cooperative and most vengeful when she punished Casey's defection. But does this reflect the imposition of sanctions per se or the effect they had on Casey's final payoff?

Table 1 shows the final payoffs for Alex and Casey that resulted from their interaction (after round 3) in Studies 1 and 2. (Although participants did not see Table 1, which compares the three responses, they always saw the final payoffs for Alex and Casey in the condition to which they were assigned; see Supplementary Materials S1 and S5.) Casey always gained by defecting, but by different amounts depending on how Alex responded. In Study 1, Alex's reputation for cooperativeness was highest when Casey gained the most by defecting (no sanctions), intermediate when Casey profited some by defecting (cooperation withdrawn), and lowest when the defection was punished. Vengefulness also tracked Casey's payoffs: Alex was seen as most vengeful when Casey's payoff was lowest and least vengeful when it was highest.

These reputational consequences could also reflect Alex's final payoffs, however, because hers were anti-correlated with Casey's ($r = -1$). Indeed, Alex profited by punishing Casey in Study 1. What would happen to Alex's reputations if punishing made her worse off than withdrawing rather than better off?

Also, why did failing to sanction lead to Alex being seen as more exploitable than punishing or withdrawing cooperation? Was it because this was the only response with a payoff of zero in Study 1, or are sanctioners seen as less exploitable regardless of their payoff from sanctioning? We address these questions in Study 2, where punishing is costly to Alex.

3. Study 2

Alex's motivation to punish was ambiguous in Study 1: Was it greed or a desire to right a wrong? By punishing Casey's failure to reciprocate, Alex inflicted a cost on a defector while also reclaiming the money she had initially given to Casey. The resulting payoff to Alex—$10—was twice the payoff Alex gained when she responded by withdrawing cooperation (Table 1). As a withdrawer, Alex kept the $5 endowment she could have given to Casey in round 3, but she did not recoup the $5 she gave to Casey in round 1.

In Study 2, we made punishment costly to Alex. Punishing still deducted $5 from Casey, but that money did not go to Alex—Alex did not recoup her initial loss by punishing. To inflict this cost in round 3, Alex had to forgo the $5 endowment she would have kept as a withdrawer. This removes greed as a possible motive for punishment.

The resulting payoffs to both parties are shown in Table 1. The final payoffs to Casey are identical to those in Study 1. But, unlike Study 1, where Casey's payoffs were negatively correlated with Alex's ($r = -1$), there was no correlation between their payoffs in Study 2 ($r = 0$). This allows us to see whether Alex's reputations for cooperativeness and vengefulness reflect payoffs to Casey or to Alex.

If Alex's reputation for cooperativeness reflects the benefits Casey gained from interacting with Alex, then they will follow the same pattern in both studies: Alex will be seen as more cooperative the higher the payoff to Casey. But if punishing tarnished Alex's reputation for cooperativeness in Study 1 because observers inferred she was motivated by greed, then her reputation for being cooperative will not suffer when she punishes in Study 2. In Study 2, Alex earns more by withdrawing cooperation than by punishing or not sanctioning.

The design of Study 2 also allows us to dissociate two possible reasons that punishing gave Alex a reputation for being less exploitable than failing to sanction in Study1. Did this inference follow from her willingness to punish per se or did it reflect the relative payoffs of punishing versus not sanctioning?

In both studies, the withdrawer's payoff from the interaction was

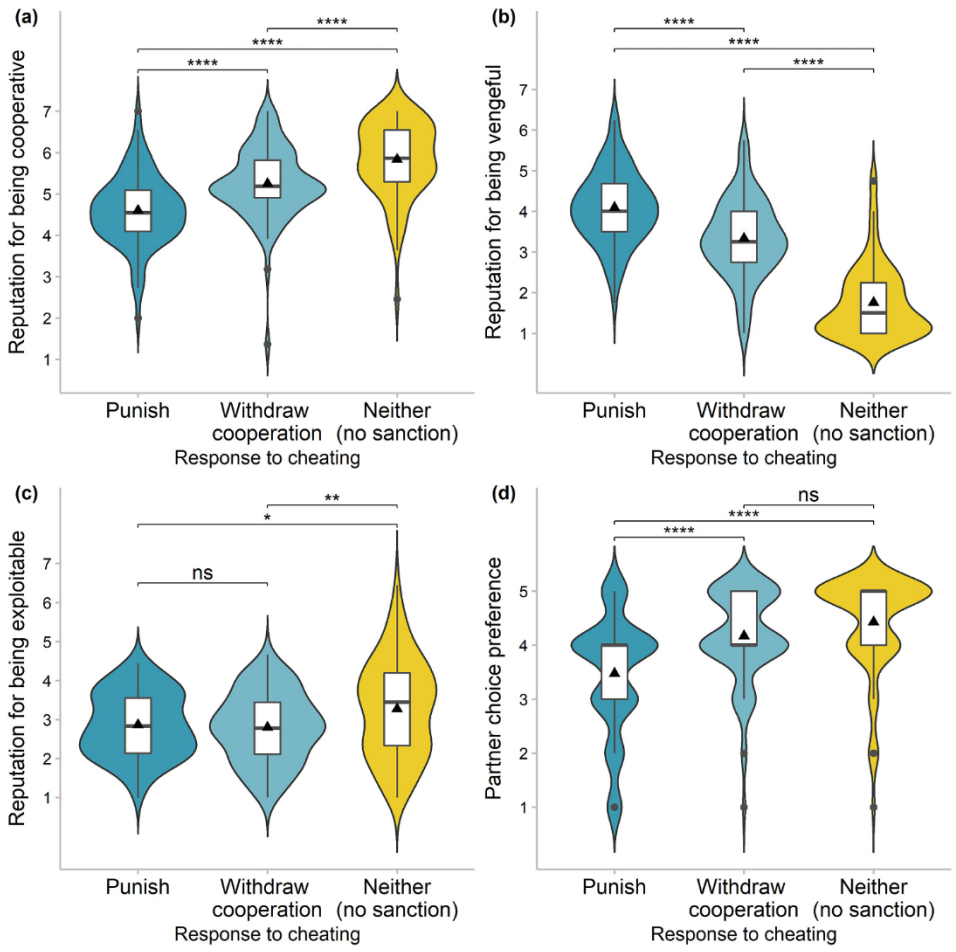

Fig. 1. Summary reputations attributed to punishers, withdrawers, and non-sanctioners: (a) cooperative reputation, (b) vengeful reputation, and (c) exploitable reputation. (d) Partner choice preferences for each responder. Boxplots show median and quartiles; Triangles represent means. *ns* > .05; * $p < .05$; ** $p < .01$; **** $p < .0001$.

Table 1
Final payoffs to the cooperator (Alex) and the defector (Casey) based on how the cooperator responded to Casey's defection. *

	Study 1 Punisher recoups initial loss		Study 2 Punisher pays, loss not recouped	
	Alex	Casey	Alex	Casey
Punish	10	5	0	5
Withdraw	5	10	5	10
No sanction	0	15	0	15

* These are payoffs *due to their interaction*; they do not count the $5 given to both parties at the beginning of each round.

positive and the no sanction payoff was zero; by contrast, punishment created a positive payoff in Study 1 and a zero payoff in Study 2. If punishing per se leads observers to see Alex as more difficult to exploit, then punishing will result in lower exploitability ratings than failing to sanction in both studies—the inference will not hinge on whether Alex's final payoff is positive versus zero. The alternative hypothesis is that inferences about exploitability are based on Alex's final payoff, regardless of her response. If earning nothing creates a reputation for being exploitable, then Alex will be seen as equally exploitable when her payoff is zero (from punishing *or* failing to sanction) and less exploitable when her payoff is positive (from withdrawing cooperation).

3.1. Methods

This study was pre-registered prior to data collection (https://osf.io/aevuh/?view_only=a1b56464033144d89d9701fbdc61ee7e) and approved by the Institutional Review Board at University of California, Santa Barbara. Data is available through the Open Science Framework (https://osf.io/yg56s/?view_only=b25462dc0b64411285e28460b02dd973). See Supplementary Material S5 for materials.

3.1.1. Participants

Participants were English speakers in the United States (n = 203,[6] 70% female, M_{age} = 19, SD_{age} = 1) recruited from an undergraduate psychology subject pool at University of California, Santa Barbara. Those who wished to participate in the study first completed a written informed consent form. The online study lasted about 8 min, and participants received a course credit for their participation.

3.1.2. Study design

The design was identical to Study 1 with two exceptions in how the

[6] The sample size in Study 2 is larger than pre-registered (n = 189), which was determined to obtain .80 power to detect the smallest effect size found in one-way ANOVA in Study 1 (Cohen's f = .229) at the .05 alpha error probability, using G*Power 3.1 (Faul et al., 2007). This is because, after the pre-registration, we decided to change the way we cluster adjectives (see footnote 2) and collected more samples to detect the smallest effect size found in the new analysis (Cohen's f = 0.221).

giver and the receiver interacted. (i) Punishment was costly (i.e., the interaction was a Dictator Game with *Reducing* Option rather than a Dictator Game with *Taking* Option). The giver had to pay $5 to reduce the receiver's earnings by $5, instead of doing this by *taking* $5 from the receiver. (ii) Instructions about the giver's options were simplified: Giving (and reducing) was all or none (no $1 increments). Givers therefore had three options in Study 2: (a) give the receiver $5, (b) give the receiver $0, or (c) pay $5 to reduce the receiver's earnings by $5.

As in Study 1, Alex gave $5 in round 1 and Casey gave $0 in round 2. In round 3, participants observed Alex respond in one of three ways (between-subjects conditions):

- Punish: Alex paid $5 to reduce Casey's earnings by $5.
- Withdraw cooperation: Alex gave $0 to Casey
- No negative sanction (keep cooperating): Alex gave $5 to Casey again.

In both studies, Alex punished by reducing Casey's earnings by $5. In Study 2, Alex had to pay $5 to accomplish this; in Study 1 Alex accomplished the same reduction by taking $5 from Casey. (See Supplementary Material S5.)

3.2. Results

3.2.1. Summary reputations

The same analysis strategy as in Study 1 was used. The factor analysis revealed a very similar three factor structure, explaining 49.5% of the total variance. Of 24 adjectives, 21 loaded on the same factors in Study 2 so, for ease of comparison, we will use the same labels for summary representations across both studies. (See Supplementary Table S3 for the factor loadings of each adjective.)

Nine adjectives, such as generous, kind, considerate, and cooperative, composed a summary reputation for being *cooperative* (Cronbach's $\alpha = .92$) Five adjectives—vengeful, aggressive, impulsive, mean, and (un)forgiving—composed a summary reputation for being *vengeful* ($\alpha = .84$). Ten adjectives, such as incompetent, unwise, exploitable, weak, and gullible, composed a summary reputation for being *exploitable* ($\alpha = .83$). The summary reputation for being *cooperative* was negatively correlated with the two others: $r\,(201) = -.63$, $p = 10^{-16}$ with *vengeful*; $-.18$, $p = .013$ with *exploitable*. The correlation between the summary reputations for *vengeful* and *exploitable* was positive in both studies, but significant only in Study 2 (.20, $p = .005$).

3.2.2. Reputational outcomes

There were significant differences in how cooperative (F [2, 200] = 37.56, $p = 10^{-14}$) and vengeful (F [2, 200] = 120.4, $p = 10^{-16}$) participants found costly punishers, withdrawers, and non-sanctioners. As in Study 1,[7] these reputations tracked the final payoffs to Casey. Alex's reputation for cooperativeness was highest when Casey gained the most by defecting (no sanctions), intermediate when Casey profited some by defecting (cooperation withdrawn), and lowest when the defection was punished (all differences significant; see Fig. 2a). Alex was seen as least vengeful when Casey's payoff was highest (no sanctions), intermediate when Casey profited some by defecting (cooperation withdrawn), and most vengeful when Casey's payoff was lowest (punished; all differences significant; see Fig. 2b).

Casey's final payoffs were uncorrelated with Alex's in Study 2: Alex's reputations for being cooperative and vengeful *did not track Alex's final payoffs*—only Casey's.

Our main results rest on the summary representations, but curious readers can consult Supplementary Table S4 for a snapshot of how people saw costly versus non-costly punishers; it compares their ratings for each adjective in Studies 1 and 2. None of the 9 adjectives contributing to the summary representations for cooperativeness in Study 2 differed for the two types of punisher. When punishing Casey's defection was costly, Alex was seen as more vengeful and impulsive than when she recouped her lost investment by punishing. (For those interested, Supplementary Tables S5 and S6 provide ratings of each adjective for withdrawers and non-sanctioners in the two studies.)

There was also a significant difference in how exploitable participants found the three responders (F [2, 200] = 9.49, $p = .0001$). The withdrawer was seen as least exploitable, with ratings lower than for the non-sanctioner (2.83 vs. 3.45, $p = 10^{-5}$) and the punisher (3.20, $p = .030$). But the exploitability of the punisher and non-sanctioner were similar ($p = .199$).

When classified by the type of response (Fig. 2c), the pattern is different from that in Study 1: Withdrawing cooperation was the only sanction that made Alex seem less exploitable in Study 2, whereas both sanctions—withdrawing and punishing—had this effect in Study 1. But when classified by Alex's final payoff due to the interaction, the results are identical across studies. Whether Alex punished or failed to sanction, a final payoff of zero led to Alex being seen as more exploitable than a final payoff that is positive. In Study 1, not sanctioning was the only response with a zero payoff for Alex; punishing and withdrawing cooperation both gave Alex a positive payoff. In Study 2, punishing and not sanctioning both led to a zero payoff for Alex; only withdrawing gave her a positive payoff.

Those curious about how costly punishers were seen compared to punishers who recouped their investment can consult Supplementary Table S4 for ratings of each adjective that loaded on *exploitability* in Study 2. The snapshot for exploitability is quite different from that for cooperativeness, where none of the 9 adjectives differed in Studies 1 and 2. Costly punishers were seen as more exploitable, unwise, incompetent, frightened, and careless than punishers who recouped their loss; they also trended toward being seen as more gullible and emotionally unstable. (*N.B.* Most of these differences were not significant when corrected for multiple [24] comparisons [Benjamini & Hochberg, 1995; Hommel, 1988].)

3.2.3. Partner choice

There were significant differences in how much participants would like to interact with the three responders (F [2, 200] = 16.76, $p = 10^{-7}$). Costly punishers were least preferred: They were rated lower than withdrawers (3.47 vs. 4.01, $p = .0007$) and non-sanctioners (4.30, $p = 10^{-7}$) (Fig. 2d). But preferences were similar for withdrawers and non-sanctioners ($p = .12$).

The partner choice preference was positively correlated with the summary reputation for being cooperative (r [201] = .72, $p = 10^{-16}$) and negatively with the ones for being vengeful ($-.44$, $p = 10^{-11}$) and exploitable ($-.21$, $p = .003$). When controlling for other reputations and which response Alex made, only the summary reputation for being cooperative ($\beta = .73$, $p < 10^{-16}$) significantly increased how much participants wanted to interact with Alex. (See Supplementary Table S7 for a full model.)

4. Conclusions

There are two ways of negatively sanctioning a defector: by withdrawing cooperation or by punishing (inflicting costs). We found that

[7] There were some differences in how Alex was evaluated in Study 1 vs. 2. Compared to Study 1, participants in Study 2 rated Alex as more exploitable (F [2, 443] = 4.26, $p = .040$) and vengeful (F [2, 443] = 4.43, $p = .036$), regardless of how Alex responded to defection (i.e., *Response* did not interact with *Study* in two-way ANOVAs). There were no differences between the two studies in how cooperative Alex appeared (F [2, 443] = 1.63, $p = .203$) or how much they would like to interact with Alex (F [2, 443] = 1.08, $p = .300$). (See also Supplementary Tables S4–6 for adjective-level comparisons.) Note, however, that our data cannot, in principle, address whether these differences are produced by the experimental manipulations (punishment was non-costly vs. costly) or characteristics of the two samples (Prolific vs. college).

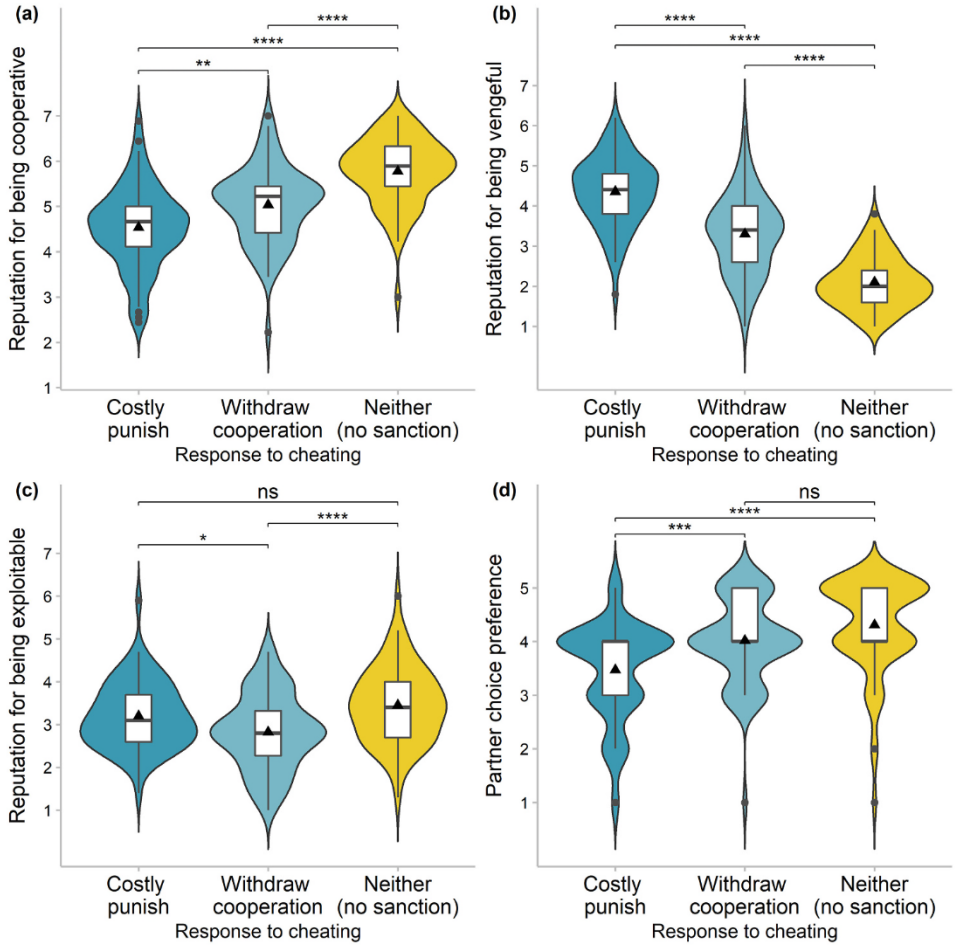

Fig. 2. Summary reputations attributed to costly punishers, withdrawers, and non-sanctioners: (a) cooperative reputation, (b) vengeful reputation, and (c) exploitable reputation. (d) Partner choice preferences for each responder. Boxplots show median and quartiles; Triangles represent means. *ns* > .05; * $p < .05$; ** $p < .01$; *** $p < .001$; **** $p < .0001$.

responding to a defector by withdrawing cooperation has better reputational consequences than inflicting punishment.

In every condition, Alex began by cooperating generously with Casey, who failed to reciprocate this generosity. But Alex's reputation varied across conditions, depending on how she sanctioned Casey's defection. As predicted, observers saw Alex as more cooperative and less vengeful when she withdrew cooperation than when she punished. They also wanted her more as a cooperation partner when she was a withdrawer than a punisher. These results did not depend on whether inflicting punishment benefitted the punisher: By punishing, Alex recouped the loss caused by Casey's defection in Study 1, but not in Study 2, where punishment was costly. Alex's reputation for being more cooperative and less vengeful perfectly tracked Casey's payoffs, but they were uncorrelated with Alex's payoffs.

Did sanctioning defections foster a reputation for being more difficult to exploit? In Study 1, both the punisher and withdrawer were evaluated as more difficult to exploit than the non-sanctioner, who continued to deliver benefits to the defecting partner. But the withdrawer was not seen as easier to exploit than the punisher, who recovered her lost investment. In Study 2, where punishment was costly, the punisher and the non-sanctioner had similar reputations for exploitability; they were both seen as easier to exploit than the withdrawer.

When classified based on response type, the exploitability results look different for Studies 1 and 2. But they are identical when classified by whether Alex's payoff from interacting with Casey was positive versus zero. Alex was always seen as more difficult to exploit when her payoff was positive. The withdrawer (positive payoff) was seen as more difficult to exploit than the non-sanctioner (zero payoff) in both studies. In Study 1, where Alex earned a positive payoff by withdrawing *or* punishing, both responses earned her a reputation as more difficult to exploit than a failure to sanction—the only response with a payoff of zero for Alex. But in Study 2, punishing at a personal cost resulted in a zero payoff to Alex, just like a failure to sanction. In this case, the punisher and non-sanctioner—both with a payoff of zero—were seen as easier to exploit than the withdrawer, whose payoff was positive.

This pattern suggests that a reputation for being difficult to exploit is inferred from a sanctioner's payoffs, rather than from punishment per se. A positive payoff always led to Alex being seen as more difficult to exploit than a payoff of zero, regardless of Alex's response to defection. How big the positive payoff was did not seem to matter—just that it was positive rather than zero.

Punishing harmed Alex's reputation for cooperativeness and her desirability as a partner. These results are consistent with theories of the evolution of reciprocity in biological markets, where partner choice is possible (Hammerstein & Noë, 2016; for converging evidence, see Arai, Tooby, & Cosmides, 2022). But they are problematic for theories that invoke group selection to explain the evolution of punishment.

A number of scholars argue that punishment is costly and thus cannot evolve without group selection (Bowles & Gintis, 2004; Fehr et al., 2002; Henrich, 2004). According to these views, individual selection cannot favor altruistic (i.e., costly) punishment because, when punishment reforms defectors, all members of the group benefit but only punishers incur a cost; that is, cooperators who do not punish are free riding on the sacrifices of punishers (the second-order free rider problem; Kiyonari &

Barclay, 2008; Panchanathan & Boyd, 2004). Because punishers have lower fitness within the group, it must be competition between groups that selects for altruistic punishment: It creates a group advantage by enforcing cooperative norms within the group. This results in more mutually beneficial transactions for members of the group, which increases that group's overall payoffs relative to groups that lack altruistic punishers.[8] Motivations to punish defectors—and to value those who do—will therefore evolve, because groups whose members are so motivated outcompete other groups. This group advantage need not come from cooperation in collective actions. Groups in which dyadic social exchange flourishes will outcompete those in which it does not, because their members will harvest more benefits from repeated, mutually beneficial cooperation.

If this explanation for altruistic punishment were correct, cooperators who punish defections would be in demand as community members and cooperative partners. Indeed, people should view punishers as *more* cooperative than non-punishers, who are free-riding on the social benefits created by punishers. The results of our studies do not support these predictions: The cooperator who punished the defector was evaluated as less cooperative than the one who did not sanction at all, and was less preferred as a future partner.

The group selection hypothesis is inconsistent with other results as well. Here we defined punishment as reducing the defector's payoff, whether the punisher recoups her loss or pays to punish. Reducing a defector's payoff should be seen as group good: Defectors who profit from their behavior not only erode cooperative norms, they cause cooperation to unravel. But the cooperator who reduced the defector's payoff was seen as *less* cooperative than the one who sanctioned by withdrawing cooperation, even though the withdrawer allowed the defector to profit from her failure to reciprocate. Moreover, a cooperator who sanctions defections—even if it is by withdrawing cooperation—should enjoy a better reputation for cooperation than a cooperator *who continues to provide benefits to the defector*. But the cooperator who did not sanction the defector at all—who instead rewarded the defector on round 3—was seen as more cooperative than even the withdrawer. The withdrawer was not preferred as a partner either: The withdrawer's partner choice ratings were similar to those for the non-sanctioner, who continued to deliver benefits to the defector. None of these results are expected on the group selection hypothesis.

Although punishing does not enhance your reputation as a cooperator, failing to impose any negative sanctions on cheaters may be costly. Non-sanctioners were seen as easier to exploit—a reputation that could attract cheaters. This finding supports the hypothesis that motivations to negatively sanction cheaters—whether by punishing or withdrawing cooperation—evolved to prevent losses by deterring mistreatment by the defector and other observers (Delton & Krasnow, 2017; Krasnow et al., 2016; Yamagishi et al., 2009). The differences between Studies 1 and 2 in perceptions of exploitability deserve further study; they suggest that sanctions will deter mistreatment more effectively when they preserve a positive payoff for the sanctioner. The efficiency of costly punishment may also matter: When their fee-to-fine ratio is >1, costly punishers may be seen as more difficult to exploit than when they fail to benefit by punishing, as in Study 2.

Withdrawing cooperation did not decrease desirability as a partner at all; this is surprising because the withdrawer stopped delivering benefits to the defector whereas the non-sanctioner continued. That difference is reflected in their reputations: Non-sanctioners were seen as more cooperative *and* more exploitable than withdrawers. Although more exploitable cooperative partners might appeal to observers with predatory intentions, exploitability was negatively correlated with partner choice in these studies. This negative evaluation of the non-sanctioner may have been offset by her very high reputation for cooperativeness, leading people to prefer her to the same degree as the (less cooperative but also less exploitable) withdrawer.

In regression analyses that included all of Alex's reputations, the only factor that continued to predict partner choice was Alex's reputation for cooperativeness. That reputation suffered most when Alex retaliated by inflicting punishment. Reluctance to punish a first-time defector might be interpreted as a forgiving strategy—a plus in repeated dyadic cooperation, considering the risk that *you* may make a mistake and defect unintentionally (Delton et al., 2012; Delton, Krasnow, Cosmides, & Tooby, 2011).

Taken together, these results highlight an advantage of withdrawing cooperation over punishment as a negative sanction: It promotes a reputation that is likely to deter exploitation while remaining favorable as a cooperative partner (see also Arai et al., 2022).

If withdrawing cooperation is better than punishment, why do people ever punish defectors? First, the benefits of being recognized as a punisher might exceed its costs in some social ecologies. When stealing resources is common in the local social ecology, as is often the case among pastoralists, acquiring a reputation for being vengeful may deter mistreatment (Cohen & Nisbett, 1996; Herrmann, Thöni, & Gächter, 2008). Second, punishers may achieve a competitive advantage over others when fee-to-fine ratio is > 1, which over-rides the reputational costs of punishing (Raihani & Bshary, 2019). Third, not all cooperative contexts have the same incentive structure; there are situations in which withdrawing cooperation is not possible (e.g., third party punishment games) or harms fellow cooperators (e.g., public goods games). Most studies of the reputational consequences of punishment used these games, and compared costly punishment to not sanctioning at all.

In third party punishment games, withdrawing cooperation is not an option for the third party, who has no opportunity to engage in cooperation with the defector. In these games, punishers were sometimes evaluated more favorably than non-sanctioners (Jordan et al., 2016; Nelissen, 2008; Raihani & Bshary, 2015b).

In public goods games, it is not possible to withdraw cooperation from a free rider without simultaneously withdrawing it from other, contributing members of the group; moreover, avoiding the free rider by leaving the group entails abandoning the benefits of group cooperation (Tooby et al., 2006). Punishment is a way of selectively sanctioning a free rider without harming other, contributing members of the group or losing the benefits of group cooperation.

Agent-based simulations show that punishing defectors in group cooperation evolves easily under many ecologically realistic conditions because, when new groups form, the defector is less likely to free ride when the punisher is also present (Krasnow et al., 2015). What evolves is a disposition to punish free riders probabilistically (not obligately). Punishing evolves because it benefits the punisher; as a side-effect, it also benefits the other cooperators in the group. Cooperators can therefore benefit from associating with occasional punishers, even if the punisher does not have a reputation for generosity. This could be why punishment in public goods games has elicited mixed results (Barclay, 2006; Kiyonari & Barclay, 2008; Mifune et al., 2020). A reputation for occasionally punishing free riders should be most attractive when groups are forming and reforming, and cooperators can choose which ones they want to join.

In summary, we demonstrated that (i) those who withdraw cooperation from cheaters are evaluated more favorably as a cooperative partner than punishers and (ii) as long as the sanction preserves a positive payoff for the sanctioner, withdrawing cooperation and inflicting punishment both protect one from acquiring a reputation that may invite exploitation.

[8] This advantage can also allow the more cooperative group to displace other groups through warfare. Note, however, that within-group cooperation and between-group violence can arise without group selection. Wrangham's (2019) domestication hypothesis, for example, does not invoke group selection. He argues that community members cooperated to execute bullies to prevent them from continuing to impose costs on them. Bullies are de facto defectors: They are men who use force to extract benefits from other community members, without providing benefits in return.

Open practices

This project was pre-registered at the Open Science Foundation prior to data collection (Study 1: https://osf.io/zwb26/?view_only=ca275e8bf4664b1381d086227bed0274; Study 2: https://osf.io/aevuh/?view_only=a1b56464033144d89d9701fbdc61ee7e).

Data availability

The data, R code, and metadata associated with this research are available at https://osf.io/yg56s/?view_only=b25462dc0b64411285e28460b02dd973.

Funding

This work was supported by an Academic Senate Grant from University of California, Santa Barbara, awarded to LC (Research Grant Account: 8586963-19900-7; https://senate.ucsb.edu/grants/faculty-research/).

Declaration of Competing Interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.evolhumbehav.2022.10.002.

References

Arai, S., Tooby, J., & Cosmides, L. (2022). Motivations to reciprocate cooperation and punish defection are calibrated by estimates of how easily others can switch partners. *PLoS One, 17*(4), Article e0267153. https://doi.org/10.1371/journal.pone.0267153

Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*(4489), 1390–1396. https://doi.org/10.1007/978-3-540-27797-2_34

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences, 111*(45), 15924–15927. https://doi.org/10.1073/pnas.1413170111

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior, 27*(5), 325–344. https://doi.org/10.1016/j.evolhumbehav.2006.01.003

Barclay, P. (2013). Strategies for cooperation in biological markets, especially for humans. *Evolution and Human Behavior, 34*(3), 164–175. https://doi.org/10.1016/j.evolhumbehav.2013.02.002

Barclay, P., & Raihani, N. (2016). Partner choice versus punishment in human Prisoner's dilemmas. *Evolution and Human Behavior, 37*(4), 263–271. https://doi.org/10.1016/j.evolhumbehav.2015.12.004

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B: Methodological, 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bone, J. E., Wallace, B., Bshary, R., & Raihani, N. J. (2016). Power asymmetries and punishment in a prisoner's dilemma with variable cooperative investment. *PLoS One, 11*(5), Article e0155773. https://doi.org/10.1371/journal.pone.0155773

Bowles, S., & Gintis, H. (2004). The evolution of strong reciprocity: Cooperation in heterogeneous populations. *Theoretical Population Biology, 65*(1), 17–28. https://doi.org/10.1016/j.tpb.2003.07.001

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

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature, 373*(6511), 209–216. https://doi.org/10.1038/373209a0

Cohen, D., & Nisbett, R. E. (1996). *Culture of honor: The psychology of violence in the South*. Westview Press Inc.

Delton, A. W., Cosmides, L., Guemo, M., Robertson, T. E., & Tooby, J. (2012). The psychosemantics of free riding: Dissecting the architecture of moral concept. *Journal of Personality and Social Psychology, 102*(6), 1252–1270. https://doi.org/10.1037/a0027026

Delton, A. W., & Krasnow, M. M. (2017). The psychology of deterrence explains why group membership matters for third-party punishment. *Evolution and Human Behavior, 38*(6), 734–743. https://doi.org/10.1016/j.evolhumbehav.2017.07.003

Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences, 108*, 13335–13340.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation and the enforcement of social norms. *Human Nature, 13*(1), 1–25. https://doi.org/10.1007/s12110-002-1012-7

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Association, 90*(4), 980–994.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*(415(6868)), 137–140.

Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature, 422*(6928), 137–140. https://doi.org/10.1038/nature01474

Funder, D. C., & Sneed, C. D. (1993). Behavioral manifestations of personality: An ecological approach to judgmental accuracy. *Journal of Personality and Social Psychology, 64*(3), 479–490. https://doi.org/10.1037/0022-3514.64.3.479

Hamann, K., Warneken, F., Greenberg, J. R., & Tomasello, M. (2011). Collaboration encourages equal sharing in children but not in chimpanzees. *Nature, 476*(7360), 328–331. https://doi.org/10.1038/nature10278

Hammerstein, P., & Noë, R. (2016). Biological trade and markets. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 371*(1687). https://doi.org/10.1098/rstb.2015.0101

Hammerstein, P., & Parker, G. A. (1982). The asymmetric war of attrition. *Journal of Theoretical Biology, 96*(4), 647–682. https://doi.org/10.1016/0022-5193(82)90235-1

Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior and Organization, 53*(1), 3–35. https://doi.org/10.1016/S0167-2681(03)00094-5

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*(5868), 1362–1367. https://doi.org/10.1126/science.1153808

Hilbe, C., & Traulsen, A. (2012). Emergence of responsible sanctions without second order free riders, antisocial punishment or spite. *Scientific Reports, 2*, 458. https://doi.org/10.1038/srep00458

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika, 75*(2), 383–386. https://doi.org/10.1093/biomet/75.2.383

Horita, Y. (2010). Punishers may be chosen as providers but not as recipients. *Letters on Evolutionary Behavioral Science, 1*(1), 6–9. https://doi.org/10.5178/lebs.2010.2

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature, 530*(7591), 473–476. https://doi.org/10.1038/nature16981

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*(4), 826–842. https://doi.org/10.1037/a0011381

Klein, S. B., Cosmides, L., Gangi, C. E., Jackson, B., Tooby, J., & Costabile, K. A. (2009). Evolution and episodic memory: An analysis and demonstration of a social function of episodic recollection. *Social Cognition, 27*(2), 283–319. https://doi.org/10.1521/soco.2009.27.2.283

Klein, S. B., Cosmides, L., Tooby, J., & Chance, S. (2002). Decisions and the evolution of memory: Multiple systems, multiple functions. *Psychological Review, 109*(2), 306–329. https://doi.org/10.1037/0033-295X.109.2.306

Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2015). Group cooperation without group selection: Modest punishment can recruit much cooperation. *PLoS One, 10*(4), Article e0124561. https://doi.org/10.1371/journal.pone.0124561

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

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy, 115*(3), 482–493. https://doi.org/10.1086/519249

Masclet, D., Noussair, C., Tucker, S., & Villeval, M.-C. (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. *American Economic Review, 93*(1), 366–380. https://doi.org/10.1257/000282803321455359

Mifune, N., Li, Y., & Okuda, N. (2020). The evaluation of second- and third-party punishers. *Letters on Evolutionary Behavioral Science, 11*(1), 6–9. https://doi.org/10.5178/lebs.2020.72

Nelissen, R. M. A. (2008). The price you pay: Cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior, 29*(4), 242–248. https://doi.org/10.1016/j.evolhumbehav.2008.01.001

Ozono, H., & Watabe, M. (2012). Reputational benefit of punishment: Comparison among the punisher, rewarder, and non-sanctioner. *Letters on Evolutionary Behavioral Science, 3*(2), 21–24. https://doi.org/10.5178/lebs.2012.22

Panchanathan, K., & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature, 432*(November), 499–502. https://doi.org/10.1038/nature02978

Przepiorka, W., & Liebe, U. (2016). Generosity is a sign of trustworthiness-the punishment of selfishness is not. *Evolution and Human Behavior, 37*(4), 255–262. https://doi.org/10.1016/j.evolhumbehav.2015.12.003

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/.

Raihani, N. J., & Bshary, R. (2015a). The reputation of punishers. *Trends in Ecology & Evolution, 30*(2), 98–103. https://doi.org/10.1016/j.tree.2014.12.003

Raihani, N. J., & Bshary, R. (2015b). Third-party punishers are rewarded, but third-party helpers even more so. *Evolution, 69*(4), 993–1003. https://doi.org/10.1111/evo

Raihani, N. J., & Bshary, R. (2019). Punishment: One tool, many uses. *Evolutionary Human Sciences, 1*, 1–26. https://doi.org/10.1017/ehs.2019.12

Revelle, W. (2021). *psych: Procedures for psychological*. Psychometric, and Personality Research.

Roberts, G., Raihani, N., Bshary, R., Manrique, H. M., Farina, A., Samu, F., & Barclay, P. (2021). The benefits of being seen to help others: Indirect reciprocity and reputation-based partner choice. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 376*(1838), 20200290. https://doi.org/10.1098/rstb.2020.0290

dos Santos, M., Rankin, D. J., & Wedekind, C. (2013). Human cooperation based on punishment reputation. *Evolution, 67*(8), 2446–2450. https://doi.org/10.1111/evo.12108

Sznycer, D., Tooby, J., Cosmides, L., Porat, R., Shalvi, S., & Halperin, E. (2016). Shame closely tracks the threat of devaluation by others, even across cultures. *Proceedings of the National Academy of Sciences, 113*(10), 2625–2630. https://doi.org/10.1073/pnas.1514699113

Tomasello, M. (2009). *Why we cooperate*. MIT press.

Tooby, J., Cosmides, L., & Price, M. E. (2006). Cognitive adaptations for n-person exchange: The evolutionary roots of organizational behavior. *Managerial and Decision Economics, 27*(2–3), 103–129. https://doi.org/10.1002/mde.1287

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*(1), 35–57.

Warneken, F., Lohse, K., Melis, A. P., & Tomasello, M. (2011). Young children share the spoils after collaboration. *Psychological Science, 22*(2), 267–273. https://doi.org/10.1177/0956797610395392

Williams, G. (1966). *Adaptation and natural selection*. Princeton University Press.

Wrangham, R. (2019). *The goodness paradox: How evolution made us both more and less violent*. Profile Books.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology, 51*(1), 110–116. https://doi.org/10.1037/0022-3514.51.1.110

Yamagishi, T., Horita, Y., Takagishi, H., Shinada, M., Tanida, S., & Cook, K. S. (2009). The private rejection of unfair offers and emotional commitment. *Proceedings of the National Academy of Sciences, 106*(28), 11520–11523. https://doi.org/10.1073/pnas.0900636106