PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Research

Cite this article: Burton-Chellew MN, Guérin C. 2021 Decoupling cooperation and punishment in humans shows that punishment is not an altruistic trait. *Proc. R. Soc. B* **288**: 20211611.
https://doi.org/10.1098/rspb.2021.1611

Received: 16 July 2021
Accepted: 18 October 2021

Subject Category:
Behaviour

Subject Areas:
behaviour, evolution

Keywords:
conditional cooperation, confused learners, public goods game, norm enforcement, social preferences, strong reciprocity

Author for correspondence:
Maxwell N. Burton-Chellew
e-mail: maxwell.burton@unil.ch

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5680340.

THE ROYAL SOCIETY PUBLISHING

Decoupling cooperation and punishment in humans shows that punishment is not an altruistic trait

Maxwell N. Burton-Chellew[1,2] and Claire Guérin[2]

[1]Department of Economics, HEC-University of Lausanne, 1015 Lausanne, Switzerland
[2]Department of Ecology and Evolution, Biophore, University of Lausanne, 1015 Lausanne, Switzerland

MNB-C, 0000-0002-5330-0379; CG, 0000-0001-7266-1631

Economic experiments have suggested that cooperative humans will altruistically match local levels of cooperation (conditional cooperation) and pay to punish non-cooperators (altruistic punishment). Evolutionary models have suggested that if altruists punish non-altruists this could favour the evolution of costly helping behaviours (cooperation) among strangers. An often-key requirement is that helping behaviours and punishing behaviours form one single conjoined trait (strong reciprocity). Previous economics experiments have provided support for the hypothesis that punishment and cooperation form one conjoined, altruistically motivated, trait. However, such a conjoined trait may be evolutionarily unstable, and previous experiments have confounded a fear of being punished with being surrounded by cooperators, two factors that could favour cooperation. Here, we experimentally decouple the fear of punishment from a cooperative environment and allow cooperation and punishment behaviour to freely separate (420 participants). We show, that if a minority of individuals is made immune to punishment, they (i) learn to stop cooperating on average despite being surrounded by high levels of cooperation, contradicting the idea of conditional cooperation and (ii) often continue to punish, 'hypocritically', showing that cooperation and punishment do not form one, altruistically motivated, linked trait.

1. Introduction

'Mistrust all in whom the urge to punish is strong!'

– F. W. Nietzsche (1885) [1]

Can human cooperation, often considered biologically unique, be explained by a phenomenon of altruistic punishment, whereby individuals punish non-cooperators for the good of society? Do people, for example, if they see someone littering in public, harming others or using public transport without paying, punish them? Economic experiments have attempted to model these situations. Key results suggest that human cooperation in such situations is reliant on the presence of 'altruistic punishers' who police a minority of non-cooperators even though they have nothing to gain [2–4].

Specifically, experiments using the repeated public goods game have shown that human cooperation is typically fragile [5–7] unless individuals can pay to inflict fines upon other individuals (to 'punish') [2,3]. A common explanation for these results is that, when punishment is not possible, cooperators retaliate against non-cooperators in the only way available to them, by reducing their contributions to the public good [8,9]. Including the option of costly punishment stabilizes contributions because cooperators are willing to pay not only to contribute, but also to punish non-cooperators, even when groups are arranged over time so that no two individuals ever interact twice (to mimic large societies where the chances of interacting again can be near zero) (altruistic punishment) [3,10–13]. Consequently, the argument goes, cooperators, no longer worried about being exploited by a minority of non-cooperators, can happily continue to contribute at high levels [14,15].

© 2021 The Author(s) Published by the Royal Society. All rights reserved.

Evolutionary models assuming these two traits, cooperation and punishment form one conjoined, altruistic, trait (dubbed 'strong reciprocity') have been proposed as explanations for the evolution of a uniquely human cooperation between non-relatives [3,10–12,16–19]. This 'Altruistic Punishment' hypothesis (also known as the 'Strong Reciprocity' hypothesis [20,21]) posits that '*Strong reciprocators bear the cost of rewarding or punishing even if they gain no individual economic benefit whatsoever from their acts.*' [14] and has been supported by experiments showing that punishment is mostly directed towards below average [3], or relatively lower [22], contributors. This suggests that altruistic cooperation and punishment are indeed correlated, leading some to suggest that governmental policies aiming to increase cooperation should aim 'to provide opportunities for the public-spirited to punish free-riders' rather than rely on institutional incentives [23,24].

However, natural selection, or individual learning of cultural traits, could often favour the decoupling of cooperation and punishment, making strong reciprocity unstable [25,26].

Instead, punishment, whereby one individual pays to harm another individual, may not be altruistically motivated [3,27–32], and previous experiments have confounded two crucial aspects. Specifically, it is not clear if individuals are continuing to contribute out of the knowledge that non-cooperators will not be able to exploit them, as has been assumed [14,15], or because they themselves are afraid of being punished. This is because everyone can be punished, and punishment is very common (around 80% of individuals punish someone). Therefore, the apparent linkage between cooperation and punishment may be an artefact of an experiment design which confounds a cooperative environment with a fear of being punished, and whereby nearly everyone both contributes and punishes.

We tested if contributing and punishing were behaviourally linked traits by replicating the seminal study of altruistic punishment [3] but with one key modification (figure 1). Instead of making all participants vulnerable to punishment ('Mutual-Punishers' scenario), we made some individuals permanently immune to punishment ('Immune-Punisher' scenario) [33]. If immune individuals are altruistically motivated, then they will behave the same as when not immune. This means they will continue to match the contributions of their group mates (conditionally cooperate [8,9,15]) and to punish lower contributors ('altruistic punishment' hypothesis).

By contrast, there is increasing evidence that participants are initially confused in public goods games, but learn from experience, and that levels of altruistic contributions have been overestimated ('confused learners' hypothesis [7]). If immune individuals are instead motivated by personal gain but require experience to learn how to play the game, then they will learn to (i) reduce their contributions despite high levels of contributions among their group mates and yet (ii) they may still punish others, even if they themselves are hypocritically contributing even less, demonstrating that cooperation (contributing) and punishment are not linked traits.

2. Methods

(a) Participants, software and location

The experiment was conducted in z-Tree and in French using publicly available instructions [9,34]. All our experimental files, data files and analysis files are freely available online [35]. We had 20 sessions each with 20 or 24 participants (420 in total) at the HEC-LABEX facility, University of Lausanne (UNIL), Switzerland. HEC-LABEX recruited participants using ORSEE and excluded all participants from previous experiments by the same authors [36]. Participants were mostly students enrolled at either UNIL or the Swiss Federal Polytechnic School (EPFL). We had a near equal gender ratio (202 female, 215 male, two other, and one declined to answer) and most of our participants were under 26 years of age (134 aged under 20, 257 aged 20–25, 23 aged 26–30, two aged 30–35, three were over 35, and one declined to answer).

(b) Experimental design and procedure

The experiment was based on a typical linear public goods game. Individuals were placed into groups of four, given 20 monetary units (1 MU = 0.04 CHF), and had to decide how many (0–20) to contribute to a group fund. All contributions were multiplied by 1.6 and then shared out equally regardless of contribution amounts. The individual return from contributing each MU (the marginal per capita return) was therefore below 1 (1.6/4 = 0.4), meaning that contributions were personally costly but beneficial for the group.

After we had explained the above public good decision mechanism (with no mention of punishment), we measured participants' social preferences using the 'strategy method'. The method measures an individual's preference for cooperating versus not-cooperating depending on how their group mates behave and is therefore used to control for an individual's beliefs about their group mates ($N = 380/420$, an experiment error meant we missed 40 participants). Specifically, it asks participants to make a separate decision for each and every possible scenario of mean contributions by their three group mates (21 scenarios from 0–20 MU). The instructions and presentation were copied from [9]. This allowed us to produce a more fine-grained analysis that controls for different putative 'social types' reported in the literature. The method categorizes individuals as either conditional cooperators (CC), who either perfectly, or at least approximately, match their group mate's mean contribution across all 21 scenarios (Pearson correlation greater than 0.5 and the amount they contribute when their group mates contribute fully is greater than their mean contribution for all 21 scenarios [37]), or free-riders (FR), who never cooperate regardless (contribute 0 MU for every possible scenario); or other/unclassified, who satisfied neither of these criteria [8,37]. Strong reciprocity theory predicts that CC will continue to cooperate even when they are immune from punishment.

We then conducted the core of the experiment. Participants knowingly play a repeated version of the same public goods game for five rounds. They were told that the group composition would change after each round, in such a way that ensured that no two individuals ever interacted twice ('perfect-stranger' matching, [2,3]). We used three different scenarios, each with either one or two distinct player roles, that were held constant for the duration of the game (figure 1). All participants in the same session faced the same scenario. The five rounds of the game therefore represent five 'one-shot' versions of the scenario, played with different people each time, and all roles held constant.

In each scenario, there were two decision phases per round of the game that followed the original design for 'altruistic punishment' [3]. In decision phase one, individuals made a simultaneous contribution to the public good, ranging from 0 to 20 MU. Then, for decision phase two, all the individuals were informed of the individual contributions and pay-offs of each of their group mates (we showed them their group mates contributions in random order on the screen). They could then, depending on their assigned role, choose to pay to deduct earnings from their group mates (to punish). The cost of punishment was 1 MU to deduct 3 MU. Individuals could spend up to 10 MU per person they punished (6 MU in sessions 1 and 2, electronic supplementary material, Methods). To make punishment equally

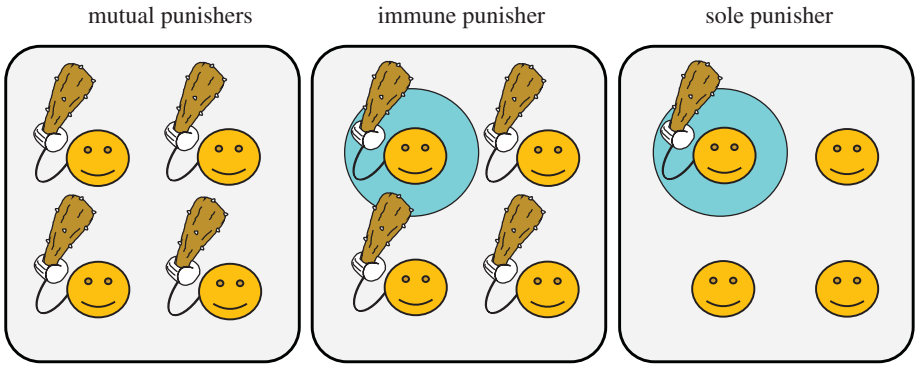

Figure 1. Experiment design. We varied either the number of group members allowed to punish others (holding a big club), or that were permanently immune to the threat of punishment (surrounded by blue shield). The Mutual-Punishers scenario replicated the original experiment design on altruistic punishment [3]. The Immune-Punisher scenario contained one group member that was immune to the threat of punishment but who could still punish group mates. The Sole-Punisher scenario contained only one group member that could punish group mates, and thus this punisher was de facto immune. (Online version in colour.)

affordable to all eligible punishers, each individual was endowed with a punishment budget of 30 MU (18 MU in sessions 1 and 2, electronic supplementary material, Methods). The most a victim of punishment (punishee) could lose was their total earnings from the contributions stage (maximum 44 MU), which meant no individuals could ever earn a negative amount from a specific round of the game. After the punishment decisions had been made, the individuals learned how many MU had been deducted from their account, i.e. how much they had been punished, but they did not know which individual(s), or how many, had punished them.

Our primary modification was to introduce individuals that were 'immune' to punishment. In the classical design, all individuals could punish all other individuals [3]. We replicated this design in our baseline scenario ('Mutual-Punishers' scenario). By contrast, in our other two scenarios, we introduced one immune individual to each group (figure 1).

In our 'Immune-Punisher' scenario, one individual, the 'immune punisher', was immune to punishment, but could still punish all the other individuals, who were not immune. This meant the immune punisher could punish up to three group mates, whereas the non-immune punishers who were analogous to a sub-group of 'mutual-punishers' in this scenario could only punish up to two group mates (figure 1). As a control, we included a scenario where the immune punisher was the sole group member that could inflict punishment ('Sole-Punisher' scenario, figure 1) [38–40], allowing us to test if immune punishers freeride on the punishment of others when they are not the sole potential punisher. The key treatment design in our experiment was whether an individual was randomly assigned to be immune or not immune, within three different background scenarios/ecologies (figure 1).

(c) Financial incentives

Each MU was worth 0.04 CHF, so 20 MU was worth 0.8 CHF (see electronic supplementary material, Methods for details of exceptions in Sessions 1–3). All earnings were rounded up to the nearest CHF, and the mean average payment was 22.60 CHF (this includes the 10 CHF showup fee) and ranged from 18 CHF to 31 CHF, with a median and a mode of 22 CHF.

(d) Statistical analyses

We analysed the data using R-Studio [41]. All statistical tests were two-tailed and the analysis code files are freely available online [35].

3. Results

When all four group members could punish (Mutual-Punishers scenario), mean contributions were stable across the five rounds between 42% and 45%, showing no significant decline over time, replicating Fehr & Gächter's original findings of stable contributions under altruistic punishment [3] ($N = 80$ individuals across four sessions; glmer: generalized linear mixed model with binomial link function controlling for session and participant; round estimate = -0.02 ± 0.019, z-value = -1.16, $p = 0.247$; figure 2; electronic supplementary material, figures S1 and S2). Median contributions remained between 8 and 9.5 out of 20 MU.

(a) Immune individuals contributed significantly less over time

By contrast, in the Immune-Punisher scenario, the immune individuals significantly decreased their mean contributions over time, from 31% to 17%, as predicted by the confused learners hypothesis and contradicting the altruistic punishment/strong reciprocity hypothesis (Immune Punisher, $N =$ 42 immune individuals across eight sessions; glmer: round estimate = -0.32 ± 0.033, z-value = -9.96, $p < 0.001$; figure 2; electronic supplementary material, figures S1 and S2). Although the non-immune individuals also significantly decreased their contributions, the decline was minor, with the mean decreasing just slightly from 41% to 35%, significantly less than the decline among the immune individuals (Immune Punisher, $N = 42$ immune and 126 non-immune individuals across eight sessions; glmer: round × immunity = -0.25 ± 0.036, z-value = -6.93, $p < 0.001$). Consequently, by the final round of the game, the differences were stark, with the median contribution for immune individuals having decreased from 5 to 0 MU, compared with from 8 to 7 MU for non-immune individuals.

A similar dynamic occurred in the Sole-Punisher scenario, where only one group member could punish, and was therefore in effect also immune. In this scenario, the immune individuals again showed a significantly greater decline in contributions, from 40% to 12%, although the sole punisher was insufficient to prevent the contributions of non-immune individuals from also declining, from 46% to 25% (sole punisher, $N = 43$ immune and 129 non-immune

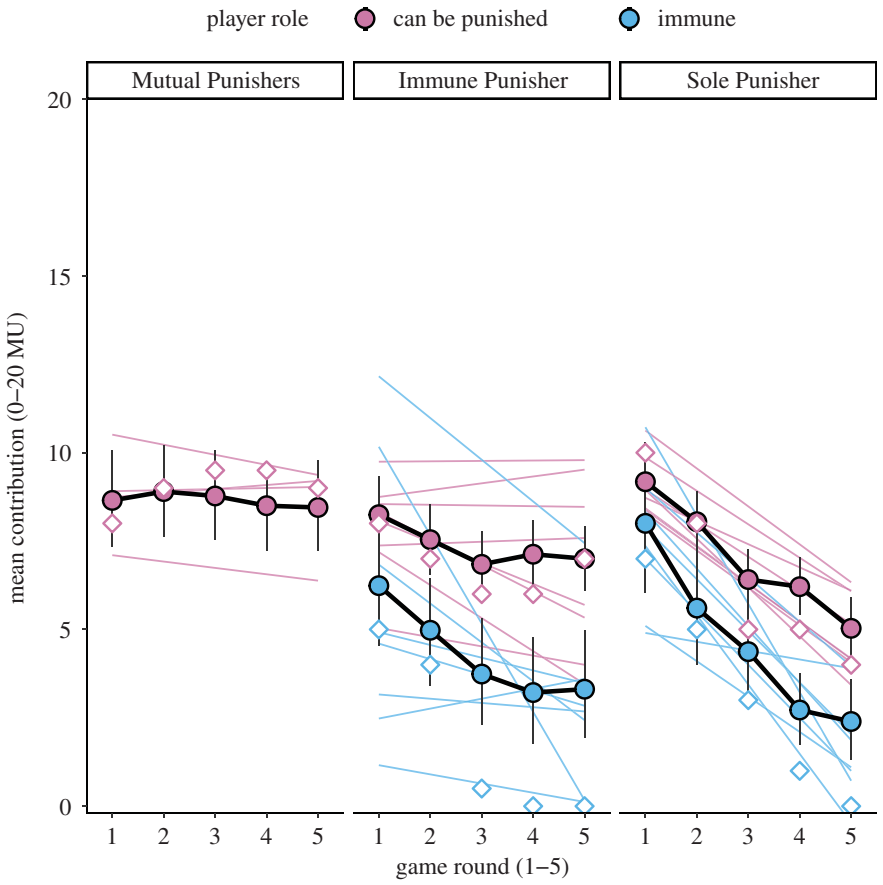

Figure 2. Immune individuals contribute less over time. Shown for each scenario are the mean (filled circles), plus 95% bootstrapped confidence interval and the median (empty diamonds) contributions per individual per round, depending on player role (magenta = non-immune; blue = immune). Mutual Punishers, $N = 80$; Immune Punisher, immune, $N = 42$; non-immune, $N = 126$; Sole Punisher, immune = 43; non-immune = 129. Coloured straight lines show, for the purposes of illustration only, linear regression estimates for each unique session × player role combination. Sessions were statistically modelled as random effects ($N = 4$ in the Mutual Punishers; eight in each of the Immune-Punisher and Sole-Punisher scenarios). Immune individuals showed a greater decline in contributions than non-immune individuals in both the Immune-Punisher scenario and the Sole-Punisher scenario.

individuals across eight sessions; glmer: round × immunity = -0.23 ± 0.035, z-value = -6.69, $p < 0.001$). Again, median contributions of immune individuals bottomed out by the final round, decreasing from 7 to 0 MU, whereas non-immune individuals were consistently higher, decreasing from 10 to 4 MU (figure 2; electronic supplementary material, figures S1 and S2).

(b) Conditional cooperators learn to not contribute

Our central finding that immune individuals learned to contribute less despite being able to punish FR and despite being surrounded by cooperators (in the Immune-Punisher scenario), still applied to 'CC'. When we repeated our above analyses, but on just those immune individuals that were classified as CC, we still found the same qualitative pattern of results, with immune CC declining significantly faster than their non-immune group mates (electronic supplementary material, Results and figures S3–S4; decline of contributions of immune CC in the Immune-Punisher scenario: from 41% to 23%, median from 8 to 1 MU, $N = 24$; in Sole-Punisher scenario: from 50% to 16%, median from 10 to 0 MU, $N = 28$). By the final round of the Immune-Punisher scenario, the immune CC had a median contribution of 1 MU despite their non-immune group mates having a median contribution of 7 MU, and despite immune and non-immune CC having started at the same level of contributions (Immune-Punisher scenario: both means = 41%, both medians = 8 MU; Sole-Punisher scenario: means = 48–50%, both medians = 10 MU, electronic supplementary material, figure S4).

These results confirm that the lower contributions of immune individuals were not simply due to them coincidentally being non-cooperators (FR) who switched strategy depending on whether they could be punished or not [13]. Instead, apparent CC, just like apparent FR, will take advantage of their immunity, not at first, but once they have learned how to play the game and will no longer match a stable, peer enforced group average (Immune-Punisher scenario) if they themselves are immune from punishment (figure 3). Crucially, our experiment design retained the cooperative environment in the Immune-Punisher scenario. This means that the decline among CC cannot be attributed to a frustration at the lack of opportunities to discipline FR or an attempt to match a declining level of contributions.

(c) Destructive punishment

Punishment was common and destructive (electronic supplementary material, figure S5, tables S1–S3, and Results). In the Mutual-Punishers scenario, which replicated the original design to test for altruistic punishment, we found that most individuals, 79% (63/80), chose to punish at least once, and 26% (21/80) chose to punish in every round (electronic supplementary material, table S2), confirming our

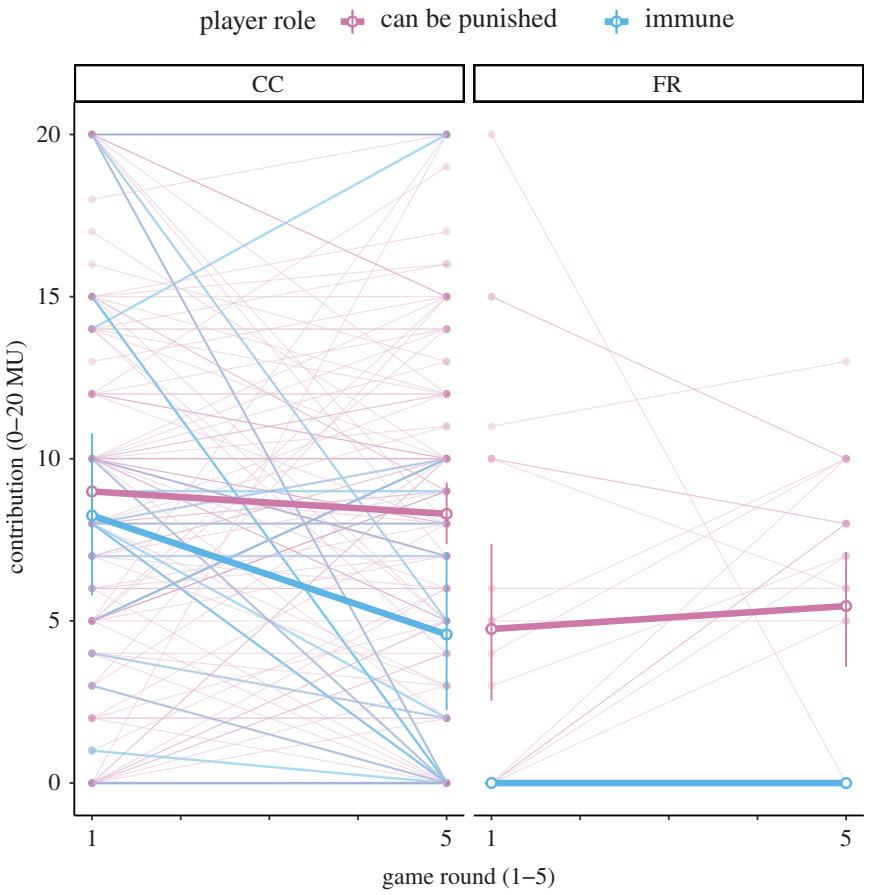

Figure 3. Social preferences and the threat of punishment. The threat of punishment converts FR into cooperators, but experience with immunity converts many CC into FR. Social preferences/type were measured before the game with the strategy method [36]. Data show the link between initial and final round contributions only. Thin lines are for each participant (filled circles), and the thick lines connect the mean contributions (empty circles) plus 95% bootstrapped confidence intervals. Magenta = non-immune (can be punished), blue = immune from punishment. The figure combines individuals from both the Mutual-Punishers and the Immune-Punisher scenarios, where there was always a threat of punishment from three group mates (or none). Sample sizes: CC, non-immune = 126; immune = 24; FR, non-immune = 24; immune = 5.

methods and participant pool were comparable with the original study where 84% of individuals punished at least once [3]. The mean average spending on punishment per round was 3.2 MU, 95% bootstrapped confidence interval = [2.74, 3.67] MU. This much punishment was ultimately destructive, with the collective costs more than outweighing the collective benefits of any increase in contributions (electronic supplementary material, Results and figure S5).

(d) Immune individuals punished less

Punishment may be motivated by other factors such as a misplaced sense of 'blind revenge' or pre-emptive strikes [2,42]. In support of this idea, we found that in the Immune-Punisher scenario, the immune punishers, who could not suffer punishment and thus should not be motivated by such vengeful motives, spent only half as much, significantly less, on punishment as the non-immune individual (figure 4; mean per round spending by: immune individuals = 1.2 MU, 95% bootstrapped confidence interval = [0.94, 1.56] MU; non-immune individuals = 2.4 MU, 95% bootstrapped confidence interval = [2.13, 2.74] MU; glmer controlling for individual and session and round of the game: immune estimate = −0.86 ± 0.328, z value = −2.62, $p = 0.009$). Immune individuals punished less in total even though they were allowed to punish more targets (three instead of just two), meaning that the immune per target spending on punishment was considerably lower (electronic supplementary material, Results).

Punishment may also be partly motivated by confusion over the costs and benefits. Supporting the idea that punishment is partly driven by confusion, we found that the frequency of punishers showed a small but significant decrease over time among the immune individuals, suggesting that individuals may not only learn to not contribute, but may also learn to not punish (electronic supplementary material, Results and figure S6).

(e) Free-riding on punishment

If punishment is altruistically motivated, then individuals will not 'freeride' on punishment. However, as our above results showed, immune individuals in the Immune-Punisher scenario punished less than the non-immune individuals, consistent with free-riding on punishment. An alternative way to investigate this question is to compare if the immune individuals were more likely to punish when they were the sole punisher and thus could not freeride on the punishing behaviours of others (Sole-Punisher scenario). This comparison between sole punishers and immune punishers also has the advantage that both types were immune from punishment and thus could not be motivated by 'revenge', although we suspect sole punishers may have felt more pressure from the experimenter to punish [43].

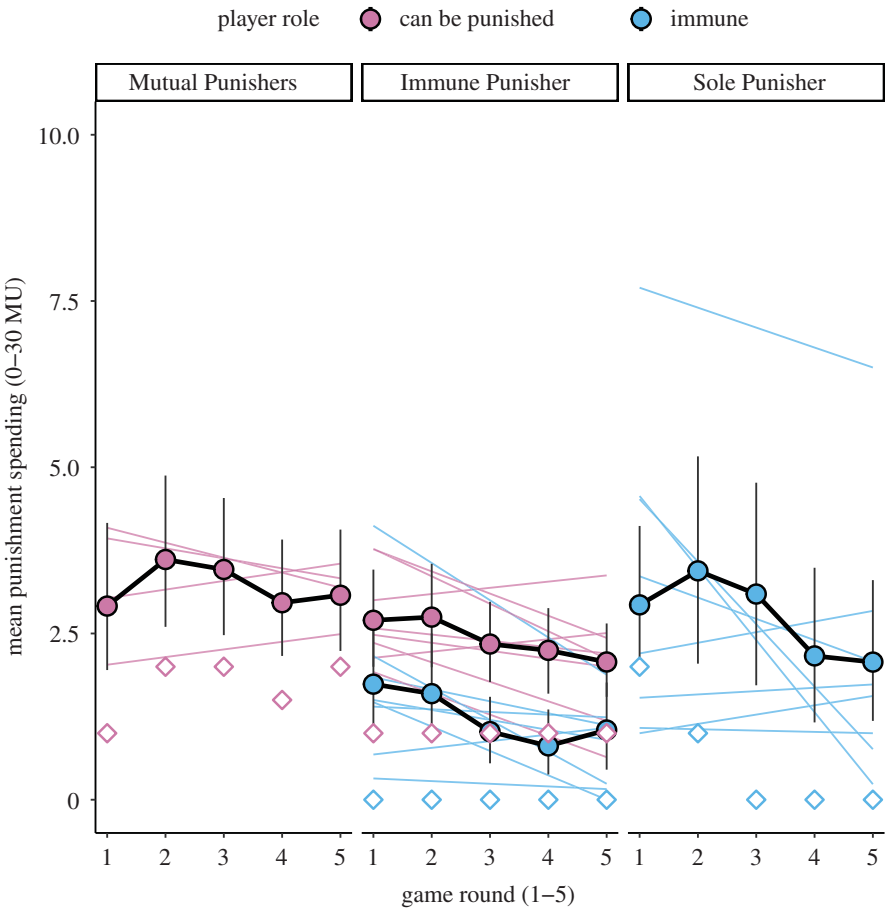

Figure 4. Immune individuals punish less. Shown for each scenario are the mean (filled circles), plus 95% bootstrapped confidence intervals, and the median (empty diamonds) mean spending per individual per round, depending on player role (magenta = non-immune; blue = immune). Mutual Punishers, $N = 80$; Immune Punisher, immune, $N = 42$; non-immune, $N = 126$; Sole Punisher, immune = 43; non-immune = 129. Coloured straight lines show, for the purposes of illustration only, linear regression estimates for each unique session × player role combination. Sessions were statistically modelled as random effects ($N = 4$ in the Mutual Punishers; eight in each of the Immune-Punisher and Sole-Punisher scenarios). In the Immune-Punisher scenario, immune individuals spent significantly less on punishment than non-immune individuals, even though they could punish more group members (three instead of just two for the non-immune).

We found that 77% of the sole punishers ($N = 33/43$) and 64% of the immune punishers ($N = 27/42$) punished at least once. The mean average spending on punishment by sole punishers was 2.7 MU, 95% bootstrapped confidence interval = [2.16, 3.34] MU, and by immune punishers was, as detailed above, 1.2 MU, 95% bootstrapped confidence interval = [0.94, 1.56] MU. However, despite the apparent difference, it is not clear if this was meaningfully more spending on punishment by the sole punishers (the bootstrapped confidence intervals did not overlap, but the glmer statistical model was non-significant: effect of scenario estimate = 0.83 ± 0.459, z-value = 1.81, $p = 0.071$).

(f) Classifying punishment behaviours

We defined the behaviour of punishers in each particular round of the game as either pro-social, anti-social or hypocritical, according to how they acted towards all their group mates ('punishment behaviours'; figure 5). This approach is more representative of behaviour than just classifying each dyadic interaction because 53% of all punishing behaviours involved an individual punishing more than one group mate ($N = 399/755$; electronic supplementary material, table S1.

In line with previous research [22,44], we found that in our Swiss participant pool, most of the punishment behaviours among non-immune individuals were pro-social and therefore appeared consistent with the altruistic punishment hypothesis. Specifically, we found that 79% of punishment behaviours in the Mutual-Punishers scenario ($N = 188/273$), and 66% of punishment behaviours among the non-immune individuals in the Immune-Punisher scenario ($N = 219/331$), could be classified as 'pro-social' (defined as exclusively against lower contributors than the punisher) (table 1).

However, among immune individuals, the instances of pro-social punishment were significantly and drastically lower, at just 49% in the Immune-Punisher scenario ($N = 38/77$), and 54.5% in the Sole-Punisher scenario ($N = 60/110$) (Fisher's exact test on counts of pro-social punishment among immune and non-immune individuals combined: $p < 0.0001$, table 1). Overall, this meant that among immune individuals, the ratio of 'pro-social' to what has previously been described as anti-social punishment was approximately 1 : 1, invalidating the view that humans can generally be described as altruistic or 'pro-social' punishers ($N = 98 : 89$, binomial sign test, two-tailed p-value = 0.559).

What could explain the high levels of anti-social (or non-pro-social) punishment among immune individuals, who

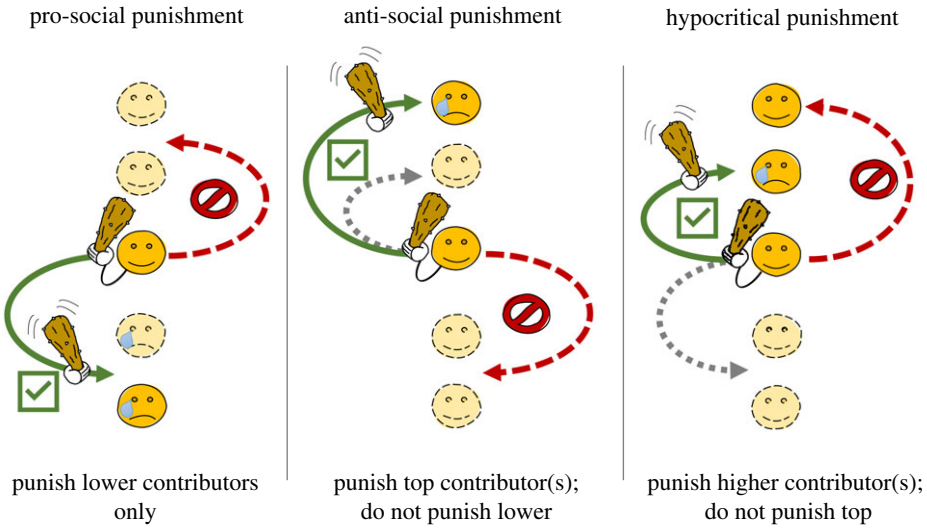

Figure 5. Classifying punishment behaviours. We classified each punisher's behaviour per round into one of three categories, or 'other/unclassified'. Our figure shows the focal individual in the middle, holding a big club, with higher/lower contributors above/below them. Punished individuals are unhappy. If the focal individual: only punished one or more lower contributors then this was 'pro-social punishment', punished the top contributor and did not punish any lower contributors, then this was 'anti-social' punishment; punished one or more equal or higher contributors, but did not punish the top contributor, then this was 'hypocritical' punishment. Green solid arrows, with a tick mark and a club, show necessary punishment for that definition; red dashed arrows with a forbidden sign, show forbidden punishment for that definition; grey dotted arrows show punishment that was allowed but not necessary for that definition. Dashed, pale individuals may or may not have existed. Solid individuals were necessary. Results are shown in table 1. (Online version in colour.)

Table 1. Hypocrisy in public goods games. A classification of punishment behaviours per individual per round (figure 5) reveals substantial hypocrisy among individuals immune from punishment. Social = the individual exclusively punished lower contributors. Anti-social = the individual punished the highest contributor(s) and did not punish socially. Hypocritical = the individual punished higher contributors, but not the highest contributor(s). Other = the behaviour did not match any of the above classifications, e.g. the individual punished both lower contributors (social), and the highest contributor (anti-social).

scenario: immunity	pro-social	anti-social	hypocritical	other	*N* obs.
Mutual Punishers: non-immune	79%	8%	8%	5%	237
Immune Punisher: non-immune	66%	14%	13%	6%	331
Immune Punisher: immune	49%	14%	34%	3%	77
Sole Punisher: immune	55%	9%	34.5%	2%	110
total: non-immune	72%	11%	11%	6%	568
total: immune	52%	11%	34%	2%	187

were based in Switzerland, a region that typically has low levels of anti-social punishment [22,44]? We found that most 'anti-social' punishment by immune individuals was hypocritical, that is, individuals punished equal or higher contributors, but not the top contributor (table 1). Taken alone, this punishment behaviour is consistent with pro-social motives; however, when combined with a lower contribution, the overall trait becomes hypocritical and communicates the message, 'do as I say, not as I do'. This challenges the view that contributions and punishment are both altruistically motivated or together form a single trait.

Finally, we analysed the consistency of each individual's punishment behaviour across all five rounds. We found that among all punishers, both immune and non-immune, only 52% restricted themselves to just pro-social punishment throughout the experiment ($N = 114$ of 220 individuals that punished) (table 2). Therefore, among individuals that punished, 48% performed at least some anti-social or hypocritical punishment at some point in the experiment. These results further invalidate the view that humans can generally be described as altruistic or 'pro-social' punishers. Instead, such exploratory behaviour is perhaps more consistent with confused learners than rational punishers.

4. Discussion

We found that we could decouple cooperation (contributions to the public good) and punishment, contradicting the altruistic punishment/strong reciprocity hypothesis. When we made one individual permanently immune to the threat of punishment, they significantly decreased their contributions (figure 2). Immune individuals also often continued to punish intermediate contributors, but not the top contributors, even though immune individuals often hypocritically contributed less (figure 5, tables 1 and 2). These results still applied to individuals previously classified as CC [8,36] (electronic supplementary material, figures S3 and S4) and show that cooperation and punishment are not linked traits, as often assumed [3,10,12,14].

Instead, our results are consistent with confused individuals initially contributing and then learning to reduce their

Table 2. Punishment consistency. Classifying the overall punishing behaviour of individuals across all five rounds.

scenario: immunity	pro-social only	never social	both[a]	*N* individuals
Mutual Punishers: non-immune	62%	6%	32%	63
Immune Punisher: non-immune[b]	51%	7%	41%	97
Immune Punisher: immune	37%	33%	30%	27
Sole Punisher: immune	48%	24%	27%	33
total: non-immune	55%	7%	38%	160
total: immune	43%	28%	28%	60
total	52%	13%	35%	220

[a]The individual was a social punisher in at least one round, and a non-social punisher in at least one round.
[b]One individual punisher was only ever classified as 'other', hence totals do not sum to 100%.

contributions [7]. Declining contributions are often attributed to frustration among impotent strong reciprocators deprived of the ability to punish, but that explanation is not possible here. This is because our immune individuals were (i) surrounded by a stable level of contributions and (ii) even had the power to punish non-cooperators. While it is possible that strong reciprocators also learn, their very existence has been inferred from previous results that required assuming individuals fully understood the consequences of their decisions [9,17].

Previous studies had concluded that humans were mostly altruistic punishers, who liked to match the cooperation of others (CC) and to punish non-cooperators ('altruistic punishment'/'strong reciprocity' hypothesis) [3,4,11,14]. Our results show that these conclusions were based on experiments that confounded a cooperative environment with a fear of punishment [3]. When an individual is free to contribute and punish as they want, they often become low contributors who punish hypocritically [39]. Punishing hypocritically increases rather than decreases inequity between individuals, challenging explanations based on inequity aversion [42,45–47].

Our results suggest that the pro-social (altruistic)/anti-social framework is perhaps a poor framework for understanding punishment behaviours (table 2). We found that many punishers, 35%, were inconsistent in their use of punishment across all five rounds, choosing to punish both pro-socially and anti-socially at different times. Overall, apparently pro-social and anti-social punishment appeared almost equally common, with 48% of all immune punishing behaviours appearing to be non-pro-social (table 1), and 48% of all punishers, across all scenarios, punishing non-pro-socially at some point (table 2). While variation in social strategies or motivations is likely ('heterogeneous preferences' [48]), we think our results demonstrate that experiments which offer participants multiple behavioural possibilities are likely to find multiple behaviours, and that there was no clear preference for altruistic punishment [29,49]. Our study only tested students based in Switzerland, so we have to be wary of over-generalizing [50,51]. However, the original findings of altruistic punishment also used students in Switzerland [3], and previous studies of anti-social punishment recorded very low levels of anti-social punishment in Switzerland [22]. Two sets of results that we contradict here using a similar participant pool.

We suggest that punishment and cooperation are social behaviours better understood through the evolutionary benefits they potentially offer to the actors [52–56]. For example, punishment can provide reputational benefits or lead to more cooperation in long-term partners [27,30,57–60]. The altruistic punishment paradigm has tested a severely restricted behavioural interaction, but outside the laboratory, behaviour is more open-ended, meaning that would-be punishers face more benefits, but also more potential costs, such as from retaliation (counter punishment) or feuds [33,61–66].

The laboratory evidence for altruistic punishment also suffers from other findings. Specifically, the fact that that costs of punishment tend to erode any collective gains from cooperation in such experiments (electronic supplementary material, figure S5), challenges the idea that altruistic punishment could even be favoured by group selection [67]. Although some evolutionary models work on the assumption that punishment will be sufficiently rare when altruistic punishers are common, meaning altruistic punishers will not be at too large a disadvantage [10]. However, our results and other experiments show that the laboratory behaviours taken as evidence for altruistic punishment in cooperative societies are frequent and easily triggered [3,22].

It is hard to rationalize hypocritical punishment. Perhaps some individuals were confused and thought they could somehow gain from the punishment, either directly, or indirectly, via a chain of interactions. The frequency of immune individuals choosing to punish significantly decreased over time, albeit it slightly (electronic supplementary material, figure S6), suggesting that individuals may be able to learn not to punish if the game were long enough. However, this is a logistical challenge in the perfect-stranger design where no two individuals can ever meet twice. Longer experiments with repeated interactions are more feasible, but they are also problematic because they can make punishment beneficial [33,58,68,69].

In conclusion, the idea that cooperation is maintained in humans by altruistic punishment appears to be a laboratory artefact of previous experiment designs, consistent with a lack of strong support from ethnographic studies [31,32,66,70,71]. Our results contradict the altruistic punishment / strong reciprocity hypothesis and support the hypothesis that many individuals are initially confused in experiments but learn from experience to contribute less (confused learners hypothesis). Immune individuals contributed less over time, and our experimental design in the Immune-Punisher scenario means this cannot be attributed to

frustration with non-cooperators. The most parsimonious explanation is that they were initially confused but learned to contribute less.

Ethics. HEC-LABEX requires that all experimental designs obtain ethical approval from the HEC-LABEX ethics committee [35]. All participants signed a consent form prior to starting.

Data accessibility. All our experimental files, data files and analysis files are freely available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.k98sf7m6r [35].

Authors' contributions. M.N.B.-C.: conceptualization, data curation, formal analysis, investigation, methodology, project administration, validation, visualization, writing-original draft, writing-review and editing; C.G.: conceptualization, investigation, methodology, project administration, writing-review and editing. All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Competing interests. We declare we have no competing interests.

Funding. We acknowledge funding from the University of Lausanne to Prof. Laurent Lehmann.

Acknowledgements. We thank Christian Thoni and Laurent Lehmann for discussion about experimental design; Laurent Lehmann and the Behaviour, Economics and Evolution seminar series at UNIL for funding, and the staff at HEC-LABEX for assistance.

References

1. Nietzsche FW. 2003 *Thus spoke Zarathustra: a book for everyone and no one*. London, UK: Penguin.
2. Fehr E, Gachter S. 2000 Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90**, 980–994. (doi:10.1257/aer.90.4.980)
3. Fehr E, Gachter S. 2002 Altruistic punishment in humans. *Nature* **415**, 137–140. (doi:10.1038/415137a)
4. de Quervain DJF *et al.* 2004 The neural basis of altruistic punishment. *Science* **305**, 1254–1258. (doi:10.1126/science.1100735)
5. Zelmer J. 2003 Linear public goods experiments: a meta-analysis. *Exp. Econ.* **6**, 299–310. (doi:10.1023/A:1026277420119)
6. Chaudhuri A. 2011 Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Exp. Econ.* **14**, 47–83. (doi:10.1007/s10683-010-9257-1)
7. Burton-Chellew MN, West SA. 2021 Payoff-based learning best explains the rate of decline in cooperation across 237 public-goods games. *Nat. Hum. Behav.* **5**, 1330–1338. (doi:10.1038/s41562-021-01107-7)
8. Fischbacher U, Gachter S, Fehr E. 2001 Are people conditionally cooperative? Evidence from a public goods experiment. *Econ. Lett.* **71**, 397–404. (doi:10.1016/S0165-1765(01)00394-9)
9. Fischbacher U, Gachter S. 2010 Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *Am. Econ. Rev.* **100**, 541–556. (doi:10.1257/aer.100.1.541)
10. Boyd R *et al.* 2003 The evolution of altruistic punishment. *Proc. Natl Acad. Sci. USA* **100**, 3531–3535. (doi:10.1073/pnas.0630443100)
11. Gintis H *et al.* 2003 Explaining altruistic behavior in humans. *Evol. Hum. Behav.* **24**, 153–172. (doi:10.1016/S1090-5138(02)00157-5)
12. Bowles S, Gintis H. 2004 The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theor. Popul. Biol.* **65**, 17–28. (doi:10.1016/j.tpb.2003.07.001)
13. Weber TO, Weisel O, Gächter S. 2018 Dispositional free riders do not free ride on punishment. *Nat. Commun.* **9**, 2390. (doi:10.1038/s41467-018-04775-8)
14. Fehr E, Fischbacher U. 2003 The nature of human altruism. *Nature* **425**, 785–791. (doi:10.1038/nature02043)
15. Fehr E, Schurtenberger I. 2018 Normative foundations of human cooperation. *Nat. Hum. Behav.* **2**, 458–468. (doi:10.1038/s41562-018-0385-5)
16. Gintis H. 2000 Strong reciprocity and human sociality. *J. Theor. Biol.* **206**, 169–179. (doi:10.1006/jtbi.2000.2111)
17. Fehr E, Henrich J. 2003 Is strong reciprocity a maladaptation? On the evolutionary foundations of human altruism, In *Genetic and cultural evolution of cooperation* (ed. P Hammerstein), pp. 55–82. Cambridge, MA: MIT Press.
18. Fowler JH. 2005 Altruistic punishment and the origin of cooperation. *Proc. Natl Acad. Sci. USA* **102**, 7047–7049. (doi:10.1073/pnas.0500938102)
19. Isler O *et al.* 2021 Contextualised strong reciprocity explains selfless cooperation despite selfish intuitions and weak social heuristics. *Sci. Rep.* **11**, 13868. (doi:10.1038/s41598-021-93412-4)
20. Carpenter J *et al.* 2009 Strong reciprocity and team production: theory and evidence. *J. Econ. Behav. Org.* **71**, 221–232. (doi:10.1016/j.jebo.2009.03.011)
21. Paál T. 2017 Altruistic punishment and strong reciprocity. In *Encyclopedia of evolutionary psychological science* (eds TK Shackelford, VA Weekes-Shackelford), pp. 1–3. Cham, Switzerland: Springer International Publishing.
22. Herrmann B, Thoni C, Gachter S. 2008 Antisocial punishment across societies. *Science* **319**, 1362–1367. (doi:10.1126/science.1153808)
23. Bowles S, Gintis H. 2002 *Homo reciprocans*. *Nature* **415**, 125–128. (doi:10.1038/415125a)
24. Bowles S, Hwang SH. 2008 Social preferences and public economics: mechanism design when social preferences depend on incentives. *J. Public Econ.* **92**, 1811–1820. (doi:10.1016/j.jpubeco.2008.03.006)
25. Lehmann L *et al.* 2007 Strong reciprocity or strong ferocity? A population genetic view of the evolution of altruistic punishment. *Am. Nat.* **170**, 21–36. (doi:10.1086/518568)
26. Rand DG, Nowak MA. 2011 The evolution of antisocial punishment in optional public goods games. *Nat. Commun.* **2**, 1–7.
27. Barclay P. 2006 Reputational benefits for altruistic punishment. *Evol. Hum. Behav.* **27**, 325–344. (doi:10.1016/j.evolhumbehav.2006.01.003)
28. Jensen K. 2010 Punishment and spite, the dark side of cooperation. *Phil. Trans. R. Soc. B* **365**, 2635–2650. (doi:10.1098/rstb.2010.0146)
29. Pedersen EJ, Kurzban R, McCullough ME. 2013 Do humans really punish altruistically? A closer look. *Proc. R. Soc. B* **280**, 20122723. (doi:10.1098/rspb.2012.2723)
30. Raihani NJ, Bshary R. 2015 The reputation of punishers. *Trends Ecol. Evol.* **30**, 98–103. (doi:10.1016/j.tree.2014.12.003)
31. Pedersen EJ, McAuliffe WHB, McCullough ME. 2018 The unresponsive avenger: more evidence that disinterested third parties do not punish altruistically. *J. Exp. Psychol.* **147**, 514–544. (doi:10.1037/xge0000410)
32. Pedersen EJ *et al.* 2020 When and why do third parties punish outside of the lab? A cross-cultural recall study. *Soc. Psychol. Pers. Sci.* **11**, 846–853. (doi:10.1177/1948550619884565)
33. Gordon DS, Puurtinen M. 2020 High cooperation and welfare despite — and because of — the threat of antisocial punishments and feuds. *J. Exp. Psychol. Gen.* **150**, 1373–1386.
34. Fischbacher U. 2007 z-Tree: Zurich toolbox for ready-made economic experiments. *Exp. Econ.* **10**, 171–178. (doi:10.1007/s10683-006-9159-4)
35. Burton-Chellew MN, Guérin C. 2021 Data from: Decoupling cooperation and punishment in humans shows that punishment is not an altruistic trait. Dryad Digital Repository. (doi:10.5061/dryad.k98sf7m6r)
36. Greiner B. 2015 Subject pool recruitment procedures: organizing experiments with ORSEE. *J. Econ. Sci. Assoc.* **1**, 114–125. (doi:10.1007/s40881-015-0004-4)
37. Thoni C, Volk S. 2018 Conditional cooperation: review and refinement. *Econ. Lett.* **171**, 37–40. (doi:10.1016/j.econlet.2018.06.022)
38. O'gorman R, Henrich J, Van Vugt M. 2009 Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. *Proc. R. Soc. B* **276**, 323–329. (doi:10.1098/rspb.2008.1082)
39. Hoeft L, Mill W. 2017 Selfish punishers: an experimental investigation of designated punishment behavior in public goods. *Econ. Lett.* **157**, 41–44. (doi:10.1016/j.econlet.2017.05.022)

40. Grieco D, Faillo M, Zarri L. 2017 Enforcing cooperation in public goods games: is one punisher enough? *J. Econ. Psychol.* **61**, 55–73. (doi:10.1016/j.joep.2017.03.007)
41. Team R. 2020 *RStudio: integrated development environment for R*. Boston, MA: RStudio, PBC. See http://www.rstudio.com.
42. Deutchman P *et al.* 2021 Punishment is strongly motivated by revenge and weakly motivated by inequity aversion. *Evol. Hum. Behav.* **42**, 12–20. (doi:10.1016/j.evolhumbehav.2020.06.001)
43. Zizzo DJ. 2010 Experimenter demand effects in economic experiments. *Exp. Econ.* **13**, 75–98. (doi:10.1007/s10683-009-9230-z)
44. Pleasant A, Barclay P. 2018 Why hate the good guy? Antisocial punishment of high cooperators is greater when people compete to be chosen. *Psychol. Sci.* **29**, 868–876. (doi:10.1177/0956797617752642)
45. Fehr E, Schmidt KM. 1999 A theory of fairness, competition, and cooperation. *Q. J. Econ.* **114**, 817–868. (doi:10.1162/003355399556151)
46. Fowler JH, Johnson T, Smirnov O. 2005 Egalitarian motive and altruistic punishment. *Nature* **433**, E1–E3. (doi:10.1038/nature03256)
47. Dawes CT *et al.* 2007 Egalitarian motives in humans. *Nature* **446**, 794–796. (doi:10.1038/nature05651)
48. Camerer CF, Fehr E. 2006 When does 'Economic Man' dominate social behavior? *Science* **311**, 47–52. (doi:10.1126/science.1110600)
49. Zhang L, Ortmann A. 2014 The effects of the take-option in dictator-game experiments: a comment on Engel's (2011) meta-study. *Exp. Econ.* **17**, 414–420. (doi:10.1007/s10683-013-9375-7)
50. Henrich J *et al.* 2006 Costly punishment across human societies. *Science* **312**, 1767–1770. (doi:10.1126/science.1127333)
51. Marlowe FW *et al.* 2008 More 'altruistic' punishment in larger societies. *Proc. R. Soc. B* **275**, 587–590. (doi:10.1098/rspb.2007.1517)
52. Clutton-Brock TH, Parker GA. 1995 Punishment in animal societies. *Nature* **373**, 209–216. (doi:10.1038/373209a0)
53. Raihani NJ, Grutter AS, Bshary R. 2010 Punishers benefit from third-party punishment in fish. *Science* **327**, 171–171. (doi:10.1126/science.1183068)
54. Raihani NJ, Thornton A, Bshary R. 2012 Punishment and cooperation in nature. *Trends Ecol. Evol.* **27**, 288–295. (doi:10.1016/j.tree.2011.12.004)
55. Kurzban R, DeScioli P. 2013 Adaptationist punishment in humans. *J. Bioecon.* **15**, 269–279. (doi:10.1007/s10818-013-9153-9)
56. Raihani NJ, Bshary R. 2015 Why humans might help strangers. *Front. Behav. Neurosci.* **9**, 39. (doi:10.3389/fnbeh.2015.00039)
57. Bshary R, Grutter AS. 2005 Punishment and partner switching cause cooperative behaviour in a cleaning mutualism. *Biol. Lett.* **1**, 396–399. (doi:10.1098/rsbl.2005.0344)
58. Gachter S, Renner E, Sefton M. 2008 The long-run benefits of punishment. *Science* **322**, 1510–1510. (doi:10.1126/science.1164744)
59. dos Santos M, Rankin DJ, Wedekind C. 2013 Human cooperation based on punishment reputation. *Evol. Int. J. Org. Evol.* **67**, 2446–2450. (doi:10.1111/evo.12108)
60. Raihani NJ, Bshary R. 2015 Third-party punishers are rewarded, but third-party helpers even more so. *Evolution* **69**, 993–1003. (doi:10.1111/evo.12637)
61. Nikiforakis N. 2008 Punishment and counter-punishment in public good games: can we really govern ourselves? *J. Public Econ.* **92**, 91–112. (doi:10.1016/j.jpubeco.2007.04.008)
62. Denant-Boemont L, Masclet D, Noussair CN. 2007 Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Econ. Theory* **33**, 145–167. (doi:10.1007/s00199-007-0212-0)
63. Nikiforakis N, Engelmann D. 2011 Altruistic punishment and the threat of feuds. *J. Econ. Behav. Org.* **78**, 319–332. (doi:10.1016/j.jebo.2011.01.017)
64. Andrighetto G *et al.* 2016 Counter-punishment, communication, and cooperation among partners. *Front. Behav. Neurosci.* **10**, 53. (doi:10.3389/fnbeh.2016.00053)
65. Balafoutas L, Nikiforakis N, Rockenbach B. 2016 Altruistic punishment does not increase with the severity of norm violations in the field. *Nat. Commun.* **7**, 13327. (doi:10.1038/ncomms13327)
66. Molho C *et al.* 2020 Direct and indirect punishment of norm violations in daily life. *Nat. Commun.* **11**, 3432. (doi:10.1038/s41467-020-17286-2)
67. Egas M, Riedl A. 2008 The economics of altruistic punishment and the maintenance of cooperation. *Proc. R. Soc. B* **275**, 871–878. (doi:10.1098/rspb.2007.1558)
68. Kiyonari T, Barclay P. 2008 Cooperation in social dilemmas: free riding may be thwarted by second-order reward rather than by punishment. *J. Pers. Soc. Psychol.* **95**, 826–842. (doi:10.1037/a0011381)
69. Rand DG *et al.* 2009 Positive interactions promote public cooperation. *Science* **325**, 1272–1275. (doi:10.1126/science.1177418)
70. Baumard N. 2010 Has punishment played a role in the evolution of cooperation? A critical review. *Mind Soc.* **9**, 171–192. (doi:10.1007/s11299-010-0079-9)
71. Guala F. 2012 Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. *Behav. Brain Sci.* **35**, 1–15.