Check for updates

Research Article

Psychological Science
1–19
© The Author(s) 2026
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09567976261439088
www.psychologicalscience.org/PS

Retributive Sentiments Track Both Deterrent and Compensatory Concerns in a Small-Scale Society and a WEIRD Sample

Léo Fitouchi[1] and Manvir Singh[2]
[1]Institute for Advanced Study in Toulouse, Department of Social and Behavioral Sciences, Toulouse School of Economics, and [2]Department of Anthropology, University of California, Davis

Abstract

Evolutionary theories suggest that retributive sentiments evolved to deter antisocial behavior, yet psychological evidence shows punitive sentiments imperfectly optimize deterrence. Recent theories resolve this paradox, suggesting that retributive sentiments serve to restore mutual benefit between partners by both compensating the victim and imposing an additional deterrent cost on the transgressor. In four experiments, we tested predictions of this hypothesis with adults from a small-scale, politically decentralized society (Mentawai horticulturalists, Indonesia; $N = 74$) and a large-scale society (the United States; $N = 600$). Consistent with compensatory concerns, participants demanded that victims be rewarded in the currency of their loss and that more severe punishment be imposed when victims suffered more, holding aggressor benefit constant. Consistent with compensation being entangled with deterrence, participants endorsed additional penalties beyond restitution that were higher for recidivism and intentional offenses. This proximate entanglement of compensation and deterrence helps resolve apparent contradictions about the function of punishment in humans.

Keywords

punishment, compensation, retributivism, justice, small-scale societies

Received 11/15/24; revision accepted 3/20/26

Humans everywhere feel outrage at social transgressions and want offenders to suffer (Black, 1993). Where do these punitive sentiments come from? According to standard evolutionary theories, punitive sentiments evolved to enforce cooperation (Gardner & West, 2004). By increasing the cost of antisocial behavior, punishment would function to deter future transgressions (Boyd, 2018; McCullough et al., 2013).

A large body of findings, however, has challenged this ultimate theory by showing that, at the proximate level, punitive sentiments often fail to track the deterrent value of punishment (Twardawski et al., 2022). If punitive sentiments function only for deterrence, for example, authorities should prescribe higher punishments when punishment is public, given that more observable punishments can deter more people from behaving badly (Molho et al., 2022). Yet people often support similar penalties when punishment is public compared with when punishment is private (Carlsmith & Darley, 2008). In fact, people even morally disapprove of punishment that would be maximally deterrent, such as harsh punishments for minor offenses (Carlsmith, 2008). Rather than seeking optimal deterrence, people are *intuitive*

Corresponding Authors:
Léo Fitouchi, Toulouse School of Economics, Department of Social and Behavioral Sciences, Institute for Advanced Study in Toulouse
Email: leo.fitouchi@iast.fr

Manvir Singh, Department of Anthropology, University of California, Davis
Email: mvrsingh@ucdavis.edu

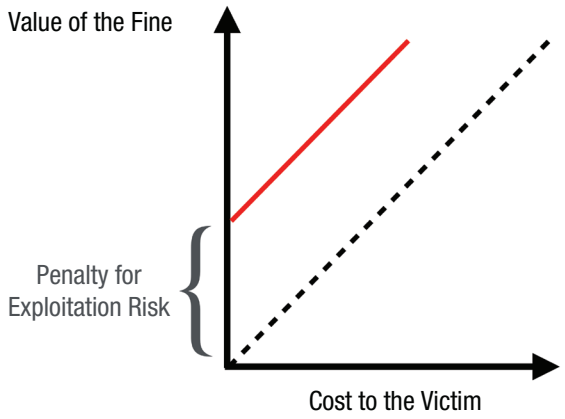

Fig. 1. The logic of restorative punishment. The value of the fine required by victims is equal to the direct cost they have suffered, plus a constant additional penalty to make up for the increased exploitation risk—both by the current offender and any other would-be offender (red line). The dotted line describes the value of a fine that would be purely restitutive—that is, that would involve no additional penalty beyond the mere replacement of the damaged resource.

retributivists: They want offenders to suffer proportionately to the harm they have done—to restore a balance, or to give offenders their "just deserts"—regardless of whether sanctioning optimally deters wrongdoing (Carlsmith & Darley, 2008; Twardawski et al., 2022).

If the proximate design of punitive sentiments is inconsistent with deterrence, should we rethink their adaptive function? Reconciling proximate design and ultimate function, an emerging view is that punishment serves a restorative function in addition to deterrence (Baumard, 2012; Fitouchi & Singh, 2023). Punishment, in this view, restores a mutually beneficial allocation of costs and benefits between the transgressor and the victim after exploitative behaviors (Fitouchi & Singh, 2023; Fitouchi et al., 2024). Although previous work has emphasized third-party interventions in the evolution of punishment (Boyd, 2018), this theory emphasizes dyadic interactions: Justice intuitions are governed by the grammar of mutual benefit; they prescribe that benefits produced by cooperation be shared in a way that benefits all partners (André et al., 2022; Levine et al., 2024). When one partner cheats the other, such as by stealing a pig, the victim is entitled to take back the benefits initially deserved (Fitouchi et al., 2024). This restoration of mutual benefit requires compensating the victim for two types of costs. The first is the direct material cost of the offense—the pig, for example, that should not have been stolen in the first place. The second is the reputational cost of the act; the offense signals to both the transgressor and other would-be offenders that the victim is the kind of person who can be easily exploited, increasing the risk of further transgressions in the future (McCullough et al., 2013). On this account, moral outrage motivates the victim to seek restoration for both costs: to demand compensation for the direct material cost of the offense and to ensure that the transgressor suffers an additional penalty to deter future exploitation (Fig. 1; André et al., 2022; Fitouchi & Singh, 2023; Fitouchi et al., 2024).

The current studies

Existing studies of intuitive retributivism have been conducted in Western, educated, industrialized, rich, and democratic (WEIRD) societies, asking participants about formal punishments, such as prison sentencing by a centralized justice system (Carlsmith & Darley, 2008; Twardawski et al., 2022). These social contexts vastly differ from the smaller-scale, politically decentralized societies where punitive sentiments have evolved (Sznycer et al., 2023). In small-scale societies more representative of human evolutionary history, punishment is administered not by a third-party state institution but by more dyadic interactions among second parties, including victims, transgressors, and their kin-based coalitions (Fitouchi & Singh, 2023). Testing hypotheses about the adaptive function of punitive sentiments, then, requires investigations in social contexts that more closely resemble this evolutionary history. Although a small number of vignette studies have included participants from small-scale societies (Barrett et al., 2016; Fessler et al., 2015), they were designed to test theories of moral judgment rather than punitive cognition.

Here, we investigate punitive cognition in a small-scale society, Mentawai horticulturalists in Indonesia (Studies 1a–b, $N = 74$, within-subject design). Despite the growing presence of the Indonesian state, much of Mentawai society continues to be politically decentralized, organized into clans, with its procedures of justice requiring the payment of resources, known as *tulou*, by the transgressor to the victim (Singh & Garfield, 2022). To test whether retributive sentiments respond to both deterrent and compensatory concerns, we presented participants with scenarios of transgressions that varied both compensation- and deterrence-relevant information. We tested two sets of predictions. The first concerns the compensatory dimension of retributive urges. If the retributive demands track compensatory concerns, we should expect the following:

1. Offenses that inflict higher direct costs on the victim should elicit greater retributive demands, independent of the benefit taken by the aggressor

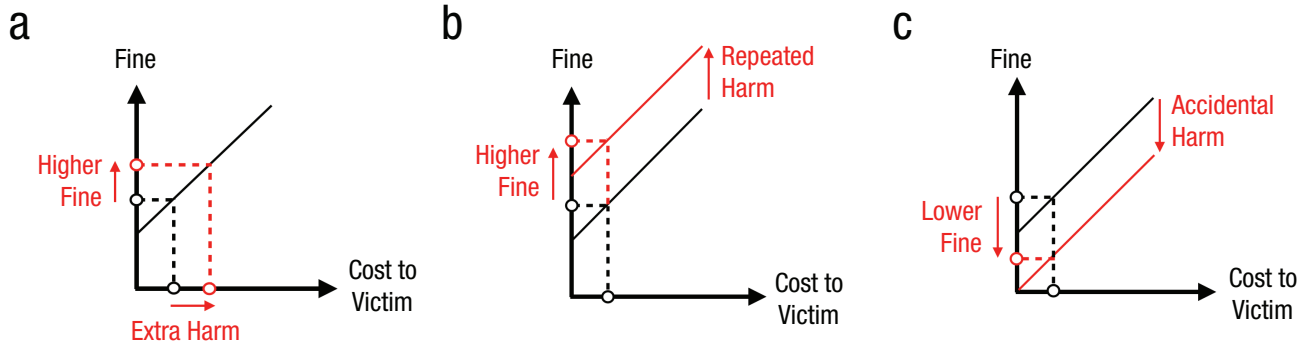

Fig. 2. Predictions of the restorative model for (a) extra-harm without extra-benefit for the transgressor, (b) repeated harm (as opposed to first-time violation; recidivism), and (c) accidental as opposed to intentional harm.

(Fig. 2a). Consider a thief who steals and consumes a victim's pig but, in doing so, deprives several piglets of their mother, leading to their deaths. Although this thief gains as much as a thief who does not kill piglets, the victim should demand a higher fine because of the higher inflicted costs. By contrast, an ultimate hypothesis based on pure deterrence predicts that retributive sentiments should be sensitive only to the benefits extracted by the offender, not to the costs imposed on victims, because the function of punishment in this view is only to signal to offenders that they can only expect a negative payoff from future aggression (McCullough et al., 2013).

2. Offenders should demand that victims be compensated in the currency of the stolen or damaged resource. Indeed, compensation requires that the value of the payment be matched as closely as possible to the loss of value suffered by the victim (Miller, 2005). Because it is easier to match the value of a stolen pig with a payment in pigs than in, say, cooking pots or coconut trees, a victim of pig theft should be repaid in pigs more often than in other currencies.

The second set of predictions concerns the deterrent aspect of retributive sentiments. If retributive sentiments track not only compensatory considerations but also the deterrence of further exploitation, we should expect the following:

3. Retributive demands should require the offender to pay an additional amount of resources beyond the direct cost of the transgression (Fig. 1).
4. When holding the direct cost of the offense constant, offenses that cause a greater reputational cost should elicit higher retributive demands. This is likely the case of repeat offenses as opposed to first-time offenses (Fig. 2b). Consider a thief who steals and eats a victim's pig but does so for the second time. Although the direct cost of the transgression is the same as for a first-time theft (the victim loses one pig in both cases), the repeat offense suggests that the punishment for their first offense was insufficient to discourage future aggression, and being exploited twice sends a stronger signal to other would-be offenders that the victim is easily exploitable, implying a greater reputational cost of repeat offenses compared with first-time offenses (Fig. 2c).
5. When holding the direct cost of the offense constant, offenses that cause a lower reputational cost should elicit lower retributive demands. This is likely the case for accidental offenses as opposed to intentional offenses (Fig. 2c). Although the direct cost of an accidental transgression is the same as if the transgression were intentional, this accidental violation is less indicative of a willingness to exploit again in the future, leading to a lower need for deterrence (Krams et al., 2013). WEIRD participants often judge accidental harms less severely than intentional ones (Young et al., 2007), yet recent theories argue for this being not a universal feature of punitive cognition but a cultural specificity of societies with low kinship intensity (Curtin et al., 2020). Because Mentawai society is more kinship-intensive than WEIRD societies, our experiments provide a strong test of the effect of intentionality on retributive judgments.

After testing those predictions with Mentawai participants, we investigated whether the punitive intuitions found in Mentawai generalized to people in a large-scale, politically centralized society. We did so by conducting a preregistered, conceptual replication of Studies 1a and 1b in a sample of 600 U.S. participants (Studies 2a and b).

Research Transparency Statement

General disclosures

Conflicts of interest: The authors declare no conflicts of interest. **Funding:** The field research was supported by a graduate research fellowship from the National Science Foundation, a Sheldon Traveling Fellowship from the Harvard Committee on General Scholarships, and a grant from the Mind, Brain, and Behavior Initiative at Harvard University. This work was also supported by France's Agence Nationale de la Recherche (under the Investissement d'Avenir program, ANR-17-EURE-0010). **Artificial intelligence:** Artificial-intelligence-assisted technologies were used only to rephrase sentences and revise some code in this research. No other artificial-intelligence-assisted technologies were used in this research or the creation of this article. **Ethics:** Studies 1a and 1b were approved by the Harvard University Committee on the Use of Human Subjects (ID No. IRB16-1080) and the Ministry of Research, Technology and Higher Education of the Republic of Indonesia. Studies 2a and 2b were approved by the Review Board for Ethical Standards in Research of the Toulouse School of Economics and Institute for Advanced Study in Toulouse. **Open Science Framework (OSF):** To ensure long-term preservation, we registered all OSF files at https://osf.io/kx4qm/.

Studies 1a and 1b disclosures

Preregistration: No aspects of Studies 1a and 1b were preregistered. **Materials:** All study materials, together with a translation codebook for the key terms used in Mentawai, are publicly available (https://osf.io/6j8m3/). **Data:** All primary data are publicly available (https://osf.io/6j8m3/). **Analysis scripts:** All analysis scripts are publicly available (https://osf.io/6j8m3/). **Computational reproducibility:** The computational reproducibility of the results in the main article (but not the Supplemental Material) has been independently confirmed by the journal's STAR team.

Studies 2a and 2b disclosures

Preregistration: All aspects of Studies 2a and 2b were preregistered prior to data collection, including design, materials, hypotheses, and analysis plan (https://osf.io/6j8m3/files/x4wmu). Note that the document was not formally preregistered; however, the time stamp indicates that it was last modified on July 25, 2025, at 16:11 UTC. Data collection began on July 25, 2025, at 16:17 UTC. There were no deviations from the preregistration. **Materials:** All study materials are publicly available (https://osf.io/6j8m3/). **Data:** All primary data are publicly available (https://osf.io/6j8m3/). **Analysis scripts:** All analysis scripts are publicly available (https://osf.io/6j8m3/). **Computational reproducibility:** The computational reproducibility of the results in the main article (but not the Supplemental Material) has been independently confirmed by the journal's STAR team.

Studies 1a and 1b

Method

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the studies.

Ethnographic context. The Siberut Mentawai are rainforest horticulturalists who reside in the river valleys of Siberut, the largest island of the Mentawai Archipelago (Indonesia; Singh et al., 2021; Tulius, 2012). They are organized into patrilineal clans known as *uma*. They traditionally resided either in longhouses (also known as *uma*) or small houses constructed nearby (Schefold, 1988; Tulius, 2012). As a result of settlement programs, most people now reside in villages that also contain clinics, schools, mosques, and churches. This research was conducted in a government village (*barasi*) and the surrounding longhouses and forest settlements in the Sarereiket region of southern Siberut.

Transgressions in Mentawai are commonly followed by a transfer of resources from offenders and their kin to victims and their kin, known as *tulou*. The practice of *tulou* predates colonial presence in Mentawai (Tulius, 2012). *Tulou* payments are typically composed of nonmonetary resources such as pigs, chickens, coconut trees, durian trees, and cooking pots (Singh & Garfield, 2022). *Tulou* payments most frequently follow the theft or damaging of another person's property (e.g., pigs, chickens, and fruit trees), as well as infractions such as adultery, spreading malicious rumors, impregnating an unmarried girl, and saying the name of a person's dead family member. Failing to share meat is notably not enforced by *tulou*, although it seems to degrade reciprocal cooperation and is believed to invite punishment from the water spirit Sikameinan (Singh et al., 2021).

Individuals demand *tulou* in several ways. In some cases, the victim's family goes directly to the offenders to demand payment and negotiate the amount. In other instances, the victim's family hires a mediator to visit the offenders and demand that they pay. The mediator travels between the parties until a resolution is reached (Singh & Garfield, 2022).

Dutch and then Indonesian administrators historically sought to reform the *tulou* system, although they were less successful in Siberut's forested interior, where this

research was conducted (Persoon, 1987). With the recent introduction of government offices, some injured parties report infractions to government officials, such as the head of the subvillage (*kepala dusun)* and the community-protection officer (*Linmas*), who are responsible for maintaining order (Singh & Garfield, 2022). Government officials then invite the offender to a meeting place and act as negotiators. Although they serve as representatives of the Indonesian government, individuals who hold these offices are drawn from the local community and do not apparently endorse Indonesian laws or standards of punishment over Mentawai customs. Instead, they function similarly to local mediators, relying on locally normative principles to determine how best to respond to infractions.

Participants. In Studies 1a and 1b, sample size was determined by convenience: We recruited as many participants as possible within the time constraints of fieldwork and of community members' busy daily activities. In Study 1a, author M.S. recruited 36 participants in several nearby villages in the Sarereiket region of Siberut Island, explaining that they would be asked about the *tulou* system. Although women attend *tulou* discussions, *tulou* payments are most often demanded and paid by men; for this reason, our sample skewed heavily male. One participant was excluded for admitting near the end of the study that she did not know *tulou* values, bringing the sample size to 35 (6 females). In Study 1b, M.S. recruited another 39 participants from the same set of nearby villages in southern Siberut. One participant did not understand the task. No data were collected from her, resulting in a sample size of 38 (16 females).

Procedures and measures. Participants in both samples were presented with nine scenarios of transgressions, in a 3 × 4, within-participant design. Each scenario was represented by an accompanying illustration, displayed in Figure 3. All scenarios involved a transgressor destroying another individual's property, such as pigs, chickens, or coconut trees. Whenever a transgression targeted livestock (i.e., pigs or chicken), it was noted that the transgressor consumed the animal at the victim's request, in accordance with local norms.

The scenarios manipulated variables capturing compensatory and deterrence concerns (control, extra-harm, recidivism, intentionality) and the type of resource stolen or damaged by the transgressor (stealing a pig, stealing a chicken, or destroying a coconut tree; Fig. 3). Of the 12 possible resource-condition combinations, three were excluded to reduce time and attention demands on participants. The remaining nine transgressions consisted of three control transgressions (stealing a pig, stealing a chicken, and destroying a coconut tree), as well as six treatment violations: two accidental transgressions (accidentally damaging a pig, accidentally damaging a coconut tree), two repeat offenses (stealing a chicken for the second time, damaging a coconut tree for the second time), and two extra-harmful ones in which the theft of a resource (a pig or chicken) resulted in the accidental death of the animal's offspring (Fig. 3; see materials for full descriptions of the transgressions read to participants).

Measure in Study 1a: prescribed value of the fine. Participants in Study 1a were asked to indicate what they considered an appropriate fine for each transgression. All participants reported fines in locally valued resources such as pigs, sago gardens, durian trees, coconut trees, cooking pots, and chickens. These represent the most common currencies for the payment of fines and bride prices in Mentawai. Community members share a general understanding of how different quantities of these resources convert into one another. At the time of data collection, a sago garden and a large iron cooking pot (size 10), for example, were both estimated to be worth approximately 1 million Indonesian rupiah (IDR). A focus group, arranged by M.S., was used to convert the value of the fines reported by participants into a common currency for our analyses. This was done at the end of data collection. Five adult men were recruited and asked to review 112 items one by one, all of them representing real-world fines in a broader study of Mentawai justice (Singh & Garfield, 2022). Men were selected because they interact more frequently with the market and are therefore more familiar with the standardized value of goods. For each item, M.S. asked: "How much is this worth in rupiah?" Group members would then discuss the item and converge on a consensus estimate (see Table S1 in the Supplemental Material available online). For many items participants achieved quick and straightforward consensus, especially for goods with clear market equivalents, such as specific types of pigs or cooking pots. The group began with these easily valued items to establish reference points and then estimated the value of less common resources in relation to these anchors (e.g., a participant would say, "Given that a garden of durian is worth about the same as a size 10 cooking pot, it should be priced similarly"). Once a consensus was reached, M.S. recorded the agreed-upon value. For 25 of the 112 goods reviewed, participants agreed only on a price range (e.g., a hen was said to be worth between 50,000 and 70,000 IDR). In those cases, we recorded the midpoint of the range as the monetary value for our analyses (60,000 IDR). These valuations were then used to convert all reported fines into a standardized monetary scale.

Measure in Study 1b: moral outrage. Participants in Study 1b ranked the nine transgressions according to the

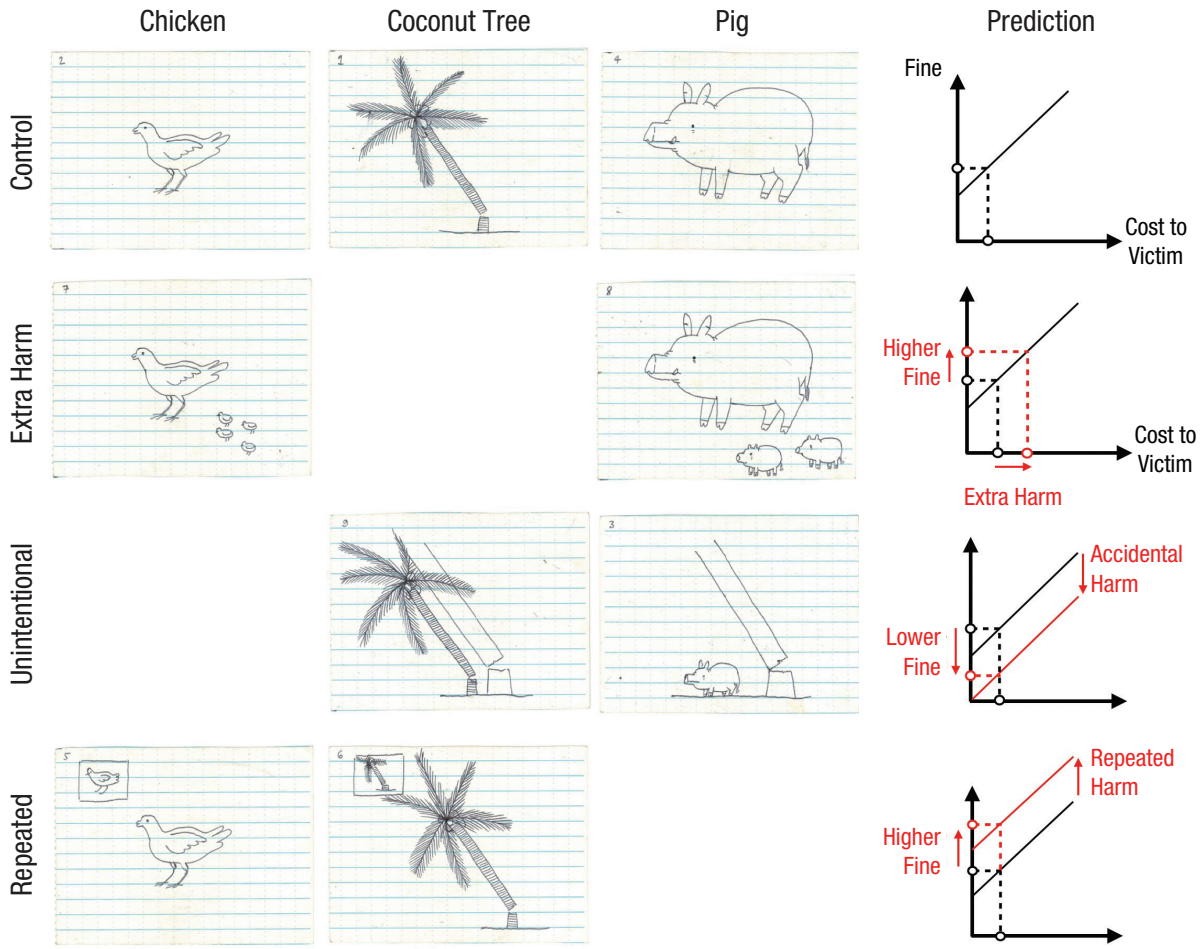

Fig. 3. The nine scenarios of transgressions, the images presented to participants to represent them, and the logic of the restorative model's prediction for each scenario.

anger they experienced in response to each act. Ordinal ranking was used instead of Likert scales, given the difficulty participants previously had using Likert scales with more than three levels. Each participant was first shown images of two randomly selected control transgressions and asked which one made them angrier. They were then presented with the third control scenario and asked whether it was more or less anger-inducing than the transgression previously ranked as most anger-inducing. For example, participants might first be shown pig killing and the felling of a coconut tree. If they indicated that pig killing elicited more anger, that image was placed to the left (denoting most anger-inducing), and the coconut-tree image was placed to the right (denoting least anger-inducing). The participants were then shown the last control scenario—chicken theft—and asked whether it was more or less anger-inducing than pig killing (previously ranked as most anger-inducing). If judged as more anger-inducing, chicken theft was placed to the left of pig killing. If it was judged as less anger-inducing, it was compared to coconut-tree felling—the scenario previously ranked as the next most anger-inducing. If it was judged as more anger-inducing than coconut-tree felling, the chicken-theft image was placed between pig killing and coconut-tree felling. Once these three control scenarios were ranked, M.S. randomly drew the remaining six treatment scenarios one at a time. For each, the participant was first asked whether the scenario elicited more anger than the scenario currently ranked as most anger-inducing. If yes, the new image was placed to the left; if not, it was sequentially compared with each lower-ranked scenario until it was either judged as more anger-inducing than another scenario or ranked as the least anger-inducing of all. This process yielded a complete ranking of all nine scenarios, ordered from most to least anger-inducing. Figure S1 in the Supplemental Material shows the distribution of anger rankings across resource-condition combinations. Thirty-eight rankings were obtained for each resource-condition combination, except for two combinations—accidental coconut-tree

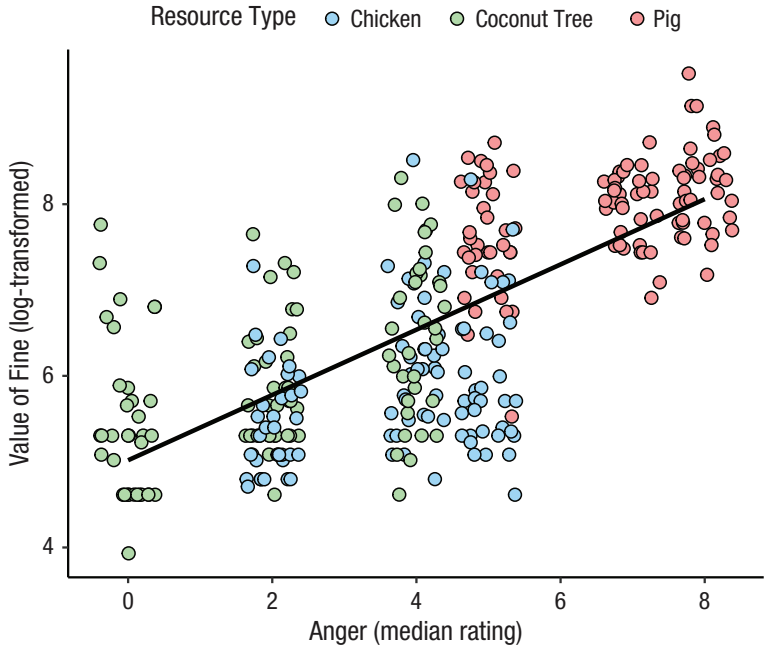

Fig. 4. Jittered scatterplot showing the median anger rating of participants in Study 1b for the nine transgressions regressed against the values of fines recommended by participants in Study 1a.

felling and accidental pig killing—which received 37 responses each because of a coding error (see Table S3 in the Supplemental Material).

Results

Moral outrage predicts fine severity. If people's demands for *tulou* payments track retributive emotions, the moral outrage expressed by participants in Study 1b should predict the *tulou* values prescribed by participants in Study 1a. To test whether this is the case, we regressed median outrage ratings of participants in Study 1b on the value of the fines prescribed by participants in Study 1a. As shown in Figure 4, participants' moral outrage at the transgressions predicted the value of recommended fines, $b = 0.38$, $SE = 0.02$, $t(276) = 21.39$, $p < .001$. This effect held when we controlled for the type of resource that was stolen or destroyed, $b = 0.23$, $SE = 0.02$, $t(274) = 9.74$, $p < .001$. These results are robust to several measures of central tendency (see Tables S3, S4 and S5 in the Supplemental Material).

Outrage and fines track cost to the victim, controlling for the benefit to the aggressor. If Mentawai participants' retributive demands track compensatory concerns, then they should track the cost suffered by the victim even when the benefit of the transgressor is held constant. To test whether this is the case, we used mixed-effects linear regressions to estimate the effect of the conditions (extra-harm, intentionality, recidivism) on both moral outrage and the fines demanded by participants. We regressed moral outrage on condition (control, extra-harm, unintentionality, repeated) and resource type (pigs, chicken, coconut tree) as fixed effects, with a random intercept for participant ID. The same predictors were used for prescribed fine value.

As shown in Figure 5, extra-harm transgressions elicited significantly higher moral outrage than control transgressions, $b = 2.12$, $SE = 0.25$, $t(334) = 8.42$, $p < .001$. The same model was used to estimate the effect of condition on the fine demanded by participants (log-transformed). As shown in Figure 5, participants prescribed significantly higher fines for extra-harm transgressions than for control transgressions, $b = 0.38$, $SE = 0.09$, $t(272) = 4.10$, $p < .001$. These results are robust to the addition of interaction terms between condition and resource type as additional predictors in both models (see Tables S7 and S8 in the Supplemental Material for full statistics).

Fines compensate victims in the currency of their loss. If Mentawai participants' retributive demands track compensatory concerns, they should not only track the direct cost suffered by the victim—regardless of benefit to transgressor—but also compensate this cost in the same currency as the original loss. To test this prediction—for example, whether victims of pig theft are more likely to

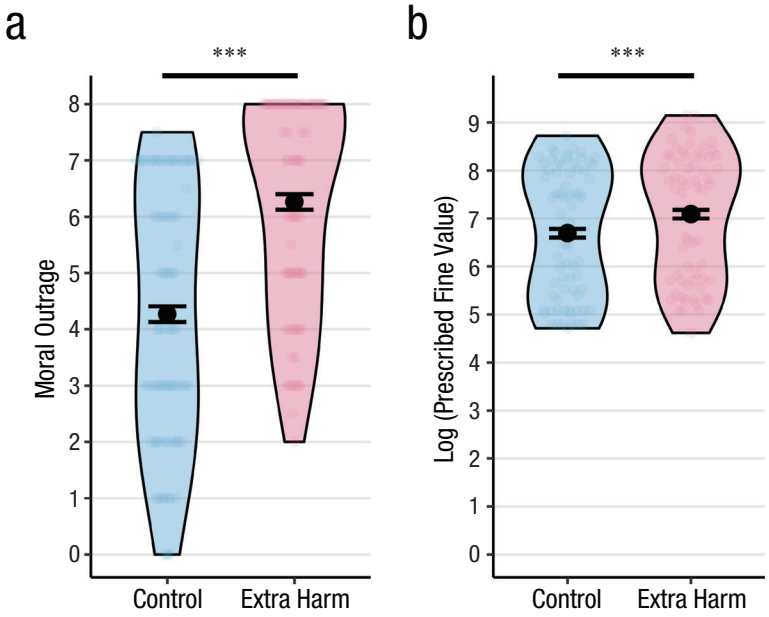

Fig. 5. Differences between the control transgressions and the transgressions that are more costly to the victim irrespective of the benefit for the aggressor (extra-harm) for (a) outrage ratings and (b) prescribed fine value (log-transformed) in Studies 1a and 1b (Mentawai sample). Error bars represent the standard errors of the least-squares means estimated from a linear mixed-effects model with condition and resource type as fixed effects and participant as a random intercept. Note that the control conditions displayed here exclude the control transgressions involving coconut trees because these did not have an extra-harm equivalent to which they could be compared; there were no extra-harm transgressions involving coconut trees in our design (see Fig. 3). The plotted difference is larger for moral outrage than for fines because fines are plotted on a log scale and outrage is measured with a ranking method. ***$p < .001$.

be repaid with pigs—we conducted separate mixed-effects logistic regressions for each resource type. Estimated fixed effects were transformed from the log-odds scale into predicted probabilities, which were then compared across resource categories.

For all resource types, fines were significantly more likely to include the specific item that had been stolen or damaged. The probability that a fine for chicken theft included chickens was 0.80, 95% confidence interval (CI) = [0.71, 0.86], significantly greater than the probability of including pigs, 0.12, 95% CI = [0.07, 0.19], $p < .001$, and coconut trees, 0.47, 95% CI = [0.37, 0.56], $p < .001$. The probability that a fine for pig theft included pigs was 0.97, 95% CI = [0.91, 0.99], significantly greater than the probability of including chickens, 0.02, 95% CI = [0.01, 0.07], $p < .001$, and coconut trees, 0.46, 95% CI = [0.34, 0.59], $p < .001$. The probability that a fine for felling a coconut tree included coconuts was 0.80, 95% CI = [0.71, 0.86], significantly greater than the probability of including chickens, 0.22, 95% CI = [0.15, 0.31], $p < .001$, and pigs, 0.23, 95% CI = [0.16, 0.32], $p < .001$. Figure 6 shows the raw frequency data.

Fines provide additional benefits to victims beyond compensating the direct costs they have incurred. If Mentawai retributive demands track not only compensatory but also deterrent concerns, they should require the offender to pay an additional amount to the victim beyond mere restitution of the direct cost suffered by the victim. To test whether this is the case, we subtracted the direct cost of the resource stolen or damaged by the transgressor from the value of fines (see Fig. 7). We then estimated the average additional payment for each resource-condition combination using least-squares means. These means were derived from a linear mixed-effects model that predicted the value of the additional payment (log-transformed) as a function of resource type (pigs, chicken, coconut trees), condition (control, extra-harm, repeated, unintentional), and their interaction, with a random intercept for participant ID.

As predicted, the additional payment required beyond pure restitution was significantly greater than zero for all resource-condition combinations (e.g., chicken control: $M = 5.06$, $SE = 0.28$, $t(211.95) = 18.23$, $p < .001$; coconut control: $M = 5.33$, $SE = 0.28$, $t(211.95) = 19.20$, $p < .001$;

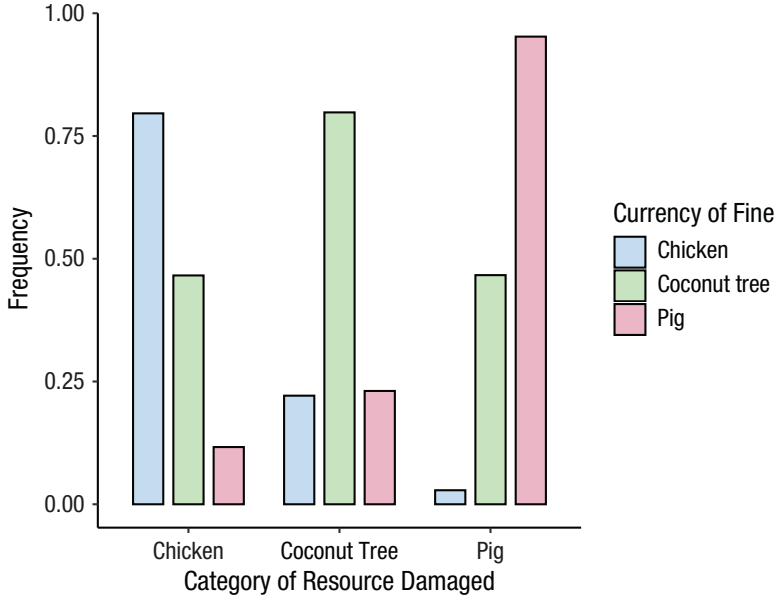

Fig. 6. The components of fines recommended by participants. *Frequency* refers to the proportions of fines recommended by participants that included a given resource. The sum of frequencies for each category of resource damaged exceeds 100% because some fines were paid with a combination of different resources.

pig control: $M = 7.51$, $SE = 0.28$, $t(211.95) = 27.04$, $p < .001$; all other ps $< .001$; see Table S9 in the Supplemental Material).

Contrary to our prediction, however, this additional payment was not equal for all types of resources. As seen in Figure 7, participants seemed to require higher additional penalties for more costly resources, instead of applying a constant additional penalty for all kinds of resources. Pairwise comparisons revealed that, in the control conditions, the mean additional payment was significantly higher for damaged pigs (a more costly resource) than for both stolen chickens (estimated mean difference: $M_{diff} = 2.44$, $SE = 0.34$, $t(269.01) = 7.11$, $p < .001$) and damaged coconut trees (estimated mean difference: $M_{diff} = 2.18$, $SE = 0.34$, $t(269.01) = 6.33$, $p < .001$). Mean additional payment was also higher for pigs than for chickens in the extra-harm conditions (estimated mean difference: $M_{diff} = 2.44$, $SE = 0.35$, $t(269.28) = 7.03$, $p < .001$) and higher for pigs than for coconut trees in the accidental conditions (estimated mean difference: $M_{diff} = 2.62$, $SE = 0.34$, $t(269.01) = 7.61$, $p < .001$). And there were no differences between chickens and coconut trees (two resources of similar value) in any of the conditions (control: $M_{diff} = -0.27$, $SE = 0.34$, $t(269.01) = -0.78$, $p = 0.997$; recidivism: $M_{diff} = -0.14$, $SE = 0.35$, $t(269.01) = -0.41$, $p = 1.00$).

Third—and consistent with our model—considerably more participants recommended fines that included only simple restitution of the lost resource, without additional payment, for *accidental* transgressions (14 of 70 fines) than for control transgressions (1 of 105), extra-harm transgressions (0 of 69), and repeat transgressions (1 of 68). An additional four fines were valued as being below the price of strict restitution. Three of them were for accidental transgressions, and one was for an extra-harm transgression.

Participants report higher outrage and fines for recidivist transgressions and lower outrage and fines for accidental transgressions. If Mentawai retributive demands track not only compensatory but also deterrent concerns, they should impose a higher additional penalty for transgressions that create a greater exploitation risk for victims, such as recidivist transgressions, and a lower additional penalty for transgressions that create a lower exploitation risk for victims, such as accidental transgressions. To estimate the effects of recidivism and intentionality on retributive demands, we used the same mixed-effects linear models used above to estimate the effect of extra harm. These models included condition (control, extra-harm, unintentional, repeated) and resource type (pigs, chicken, coconut trees) as fixed effects, with a random intercept for participant ID. These predictors were used to predict both moral outrage and prescribed fine value (log-transformed).

As shown in Figure 8, repeat transgressions elicited significantly higher outrage, $b = 1.76$, $SE = 0.25$, $t(334) = 7.02$, $p < .001$, and significantly higher fines, $b = 0.66$, $SE = 0.09$, $t(272.52) = 7.10$, $p < .001$, compared with first-time

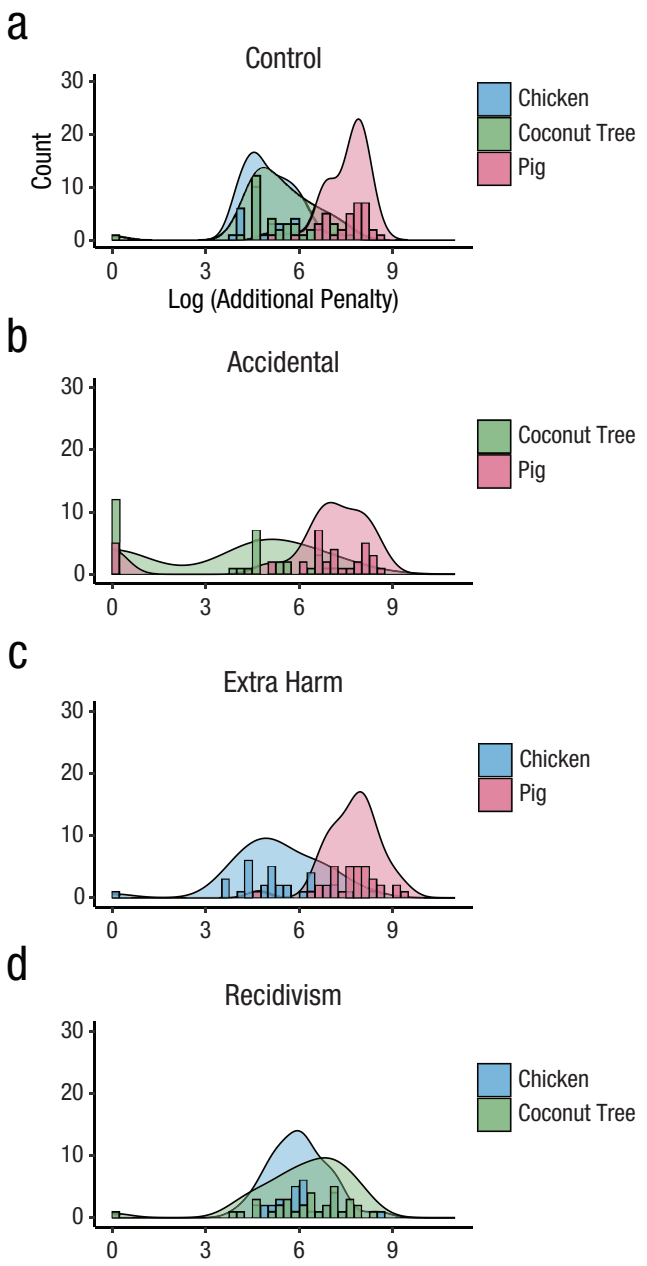

Fig. 7. Distributions of payments prescribed beyond restitution of the direct cost of the lost resource in Study 1a. In (a) we show these additional penalties for control transgressions, in (b) accidental transgressions, in (c) extra-harm transgressions, and in (d) repeat transgressions. A value of 0 means that the fine asked only for strict restitution of the lost resource—that is, that its value was equal to the value of the resource lost by the victims. Three fines for accidental offenses were lower than the value of the resource damage; these have been included here as 0.

transgressions. As shown in Figure 9, accidental transgressions elicited significantly lower outrage, $b = -1.65$, $SE = 0.25$, $t(334) = -6.51$, $p < .001$, and significantly lower fines, $b = -0.36$, $SE = 0.09$, $t(272.08) = -3.86$, $p < .001$, compared with intentional transgressions. These results were robust to the addition of interaction terms between condition and resource type as additional predictors in both models (see Tables S7 and S8 for full statistics).

Fines track the predicted reputational cost of transgressions. If participants make retributive demands to receive compensation for both direct costs and reputational costs, then reputational costs should drive variation in fine values over and above the direct costs of transgressions. Specifically, our model predicts that, after compensating for direct costs, retributive demands should be highest for recidivist transgressions (as they impose the greatest reputational costs) and lowest for unintentional offenses (as they impose the least reputational costs); extra-harm transgressions should be equivalent to those in the control condition after compensating for the direct cost, because the reputational cost is kept the same across both treatments (Fig. 2).

To test these predictions, we regressed the prescribed fine value (log-transformed) on resource type (pigs, chickens, coconut trees), condition (control, extra-harm, unintentional, repeated), and the direct cost of the transgression (in log Indonesian rupiah [IDR]), with a random intercept for participant ID. As predicted, after accounting for the compensation of direct costs, participants demanded significantly higher fines for recidivist transgressions than for control transgressions, $b = 0.69$, $SE = 0.10$, $t(271) = 7.25$, $p < .001$, significantly lower fines for unintentional transgressions, $b = -0.38$, $SE = 0.09$, $t(271) = -4.07$, $p < .001$, and fines no different from the control condition for extra-harm transgressions, $b = -0.03$, $SE = 0.26$, $t(271) = 0.10$, $p = .92$. These results were robust when we included an interaction term between condition and resource type in the model (see Table S11).

Studies 2a and 2b

Studies 1a and 1b provided evidence that, in a small-scale society without centralized justice institutions, people's retributive sentiments track parameters consistent with the restorative model. However, our sample size was small given the difficulty of acquiring experimental data in these settings. To test the robustness of our results and their generalizability across cultures, we conducted a preregistered conceptual replication of Studies 1a and 1b with a larger sample of U.S. participants (Studies 2a and 2b).

Method

Participants. A sample of 602 U.S. participants (Study 2a: $N = 300$; 149 males, 151 females; $M_{age} = 45.73$, $SD_{age} = 12.96$; Study 2b: $N = 302$; 149 males, 153 females; $M_{age} = 44.02$, $SD_{age} = 13.54$) were recruited from Prolific. Our target sample size (300 for each study) was determined

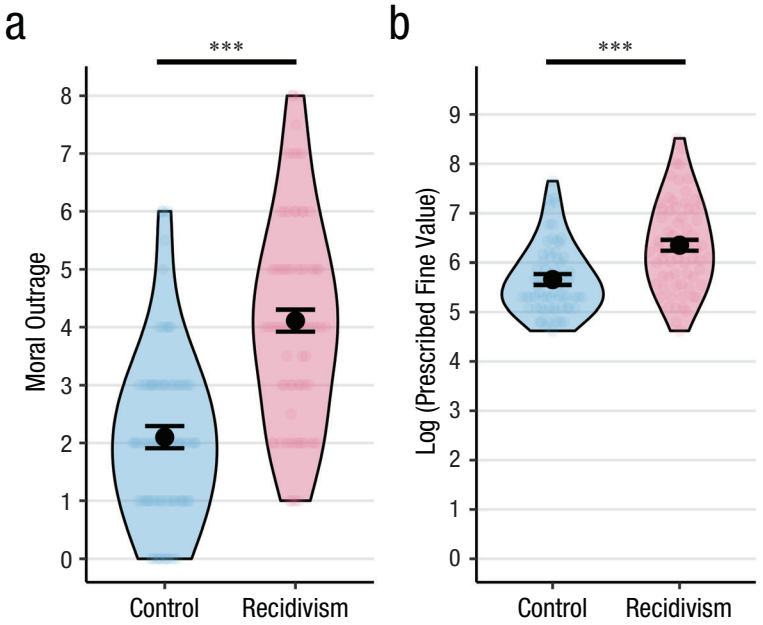

Fig. 8. Differences between the control transgressions and the recidivist transgressions for (a) outrage ratings and (b) prescribed fine value (log-transformed) in Studies 1a and 1b (Mentawai sample). Error bars represent the standard errors of the least-squares means estimated from a linear mixed-effects model with condition and resource as fixed effects and participant as a random intercept. Note that the control conditions displayed here exclude the control transgressions involving pigs because these did not have a recidivist equivalent to which they could be compared; there were no recidivist transgressions involving pigs in our design (see Fig. 3). The plotted difference is larger for moral outrage than for fines because fines are plotted on a log scale and outrage is measured with a ranking method. ***$p < .001$.

by priori power analysis using G*Power 3 (Faul et al., 2007). To conservatively estimate the power required, we approximated our most power-hungry analysis as a set of nine dependent *t* tests—one for each comparison between each control condition and the three treatment conditions across the three resource types. With a Bonferroni correction for nine tests ($\alpha = .005$), 296 participants are required to detect a small effect ($d_z = 0.20$) with 80% power. Our target sample size thus provided better than 80% power for our most power-hungry analyses. Consistent with our preregistration, participants who failed the attention check (10 in Study 2a; 15 in Study 2b) were excluded, bringing our sample size to 290 for Study 2a and 287 for Study 2b.

Procedures and measures. After consenting, participants were presented with 9 scenarios in a 3 × 4 within-subjects design. The scenarios were designed to mirror the Mentawai scenarios as closely as possible by preserving the same structure of transgressions while changing only the local context to ensure ecological validity for U.S. participants. The scenarios described the theft and killing of farm animals among two neighboring farmers, Chris and Alex, who live in rural areas of the United States. We changed the types of resources damaged (cows, hogs, and vegetable gardens) both to maintain ecological validity and to have resources of higher absolute value, given the much higher purchasing power of people in the United States compared to people in Mentawai.

As in Studies 1a and 1b, the scenarios manipulated variables capturing compensatory and deterrence concerns (control, extra-harm, recidivism, intentionality) and the type of resource stolen or damaged by the transgressor (killing a cow, stealing a hog, damaging a vegetable garden; see Table S12 in the Supplemental Material for full descriptions of the nine scenarios). As in Studies 1a and 1b, three of the 12 possible resource-condition combinations were excluded. The remaining transgressions consisted of three control transgressions (stealing a hog, killing a cow, stealing all the vegetables from the neighbor's vegetable garden) as well as six treatment transgressions: two accidental transgressions (accidentally killing a cow, accidentally damaging a vegetable garden), two repeat transgressions (stealing a hog for the second time, stealing from the vegetable garden for a second time), and two extra-harmful transgressions in which the theft or killing of a resource (a hog or a cow) resulted in the accidental death of the animal's offspring (see Table S12).

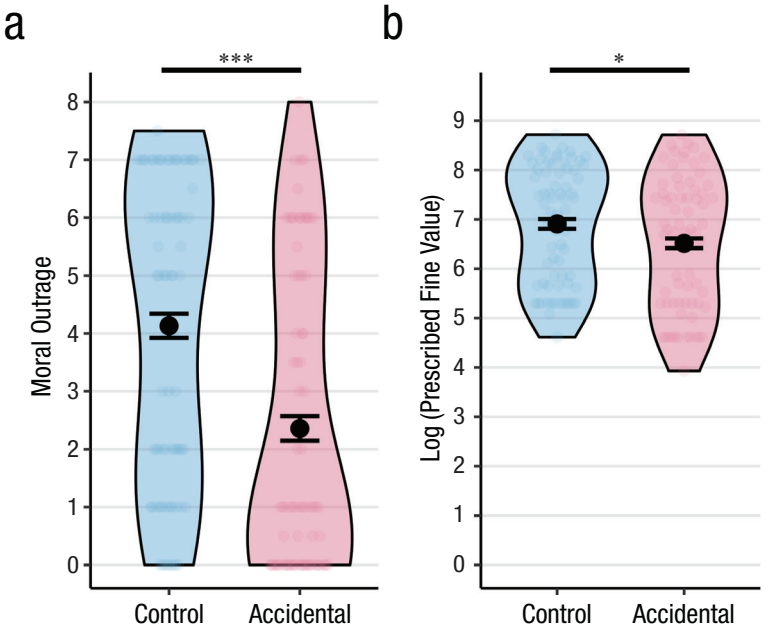

Fig. 9. Differences between the control transgressions and the unintentional transgressions for (a) moral-outrage ratings and (b) prescribed fine value (log-transformed) in Studies 1a and 1b (Mentawai sample). Error bars represent the standard errors of the least-squares means estimated from a linear mixed-effects model, with condition and resource as fixed effects and participant as a random intercept. Note that the control conditions displayed here exclude the control transgressions involving chickens because these did not have an accidental equivalent to which they could be compared; there were no accidental transgressions involving chickens in our design (see Fig. 3). The plotted difference is larger for outrage than for fines because fines are plotted on a log scale, and outrage is measured with a ranking method. *$p < .05$. ***$p < .001$.

Because the value of these resources varies more in market-integrated societies like the United States, each vignette specified the value of the resource stolen or damaged in dollars ($500 for adult hogs; $1,500 for adult cows; $100 for damage to the vegetable garden). We did so to ensure that participants had a common understanding of the value of the resource at hand.

Measure in Study 2a: Prescribed value of the fine. Unlike with Mentawai participants, who answered fine values in nonmonetary resources such as pigs, chickens, or cooking pots, U.S. participants were asked to provide fine values in U.S. dollars. This was to ensure ecological validity in a society where, unlike among the Mentawai, most exchanges use money. Participants were asked, "How much money (in $) should Chris pay to Alex to make things right?" This was a free-response question for which participants could type any number. To ensure accuracy and prevent typos, we then asked participants to retype their response in another field; they could only proceed to the next question if the two entries matched. Note that because all fines were expressed in the same currency, this study could not examine whether participants prescribed fines in the same currency as the victim's loss, as we did in Study 1a.

Measure in Study 2b: Moral outrage. Unlike Mentawai participants, who struggled with Likert scales (leading us to prefer a ranking method), U.S. participants completed a two-item scale asking, "To what extent does this act make you angry/outraged?" (0 = *not at all*, 4 = *moderately*, 6 = *extremely*). These questions showed high internal consistency and were averaged ($\alpha = .97$).

Results

Moral outrage predicts fine severity. Consistent with our preregistration, we regressed mean anger ratings of participants in Study 2b on the (log-transformed) value of the fines prescribed by participants in Study 2a. Replicating Studies 1a and 1b, the moral outrage of participants in Study 2b predicted the value of the fines recommended by participants in Study 2a, $b = 0.70$, $SE = 0.02$, $t(2319) = 35.76$, $p < .001$. This effect held when we controlled for

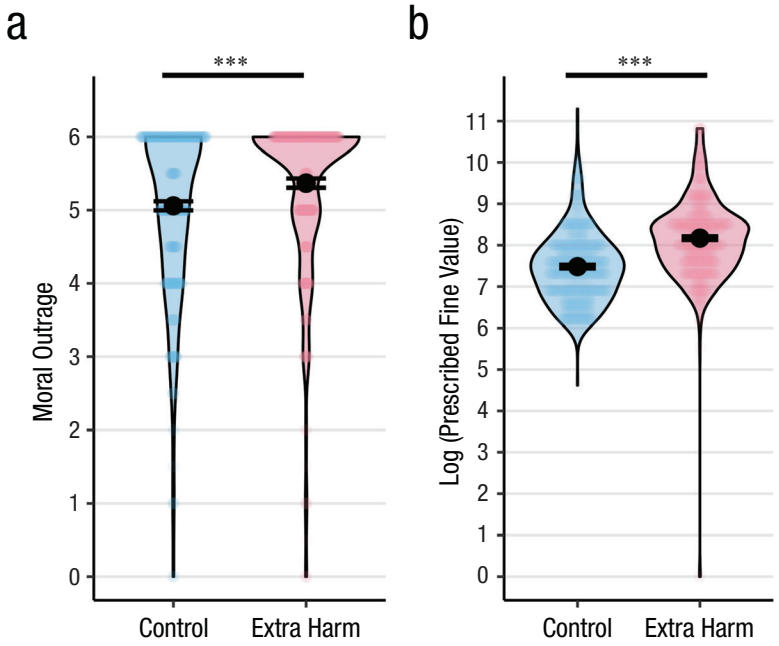

Fig. 10. Differences between the control transgressions and the extra-harm transgressions for (a) moral-outrage ratings and (b) prescribed fine value (log-transformed) in Studies 2a and 2b (U.S. participants). Error bars represent the standard errors of the least-squares means estimated from a linear mixed-effects model, with condition and resource as fixed effects and participant as a random intercept. Note that the control conditions here exclude the control transgressions involving gardens because these did not have an extra-harm equivalent to which they could be compared; there were no extra-harm transgressions involving gardens in our design (see the Method section). ***$p < .001$.

resource type, $b = 0.38$, $SE = 0.01$, $t(2317) = 26.72$, $p < .001$, and was robust to alternative measures of central tendency (see Table S13 in the Supplemental Material).

Outrage and fines track cost to the victim, controlling for the benefit to the aggressor. Consistent with our preregistration, we regressed moral outrage on condition (control, extra-harm, accidental, repeated) and resource type (hog, cow, vegetables) as fixed effects, with a random intercept for participant ID. Replicating Studies 1a and 1b, extra-harm transgressions elicited significantly higher outrage than control transgressions, $b = 0.34$, $SE = 0.05$, $t(2291) = 6.54$, $p < .001$ (see Fig. 10). As preregistered, the same model was used to estimate the effect of condition on the fines prescribed by participants (log-transformed). As shown in Figure 10, participants prescribed significantly higher fines for extra-harm transgressions than for control transgressions, $b = 0.69$, $SE = 0.03$, $t(2315) = 20.66$, $p < .001$. Exploratory analyses revealed that these results were robust when we included interaction terms between condition and resource type as additional predictors in both models (see Tables S14 and S15 in the Supplemental Material).

Fines provide additional benefits to victims beyond compensating the direct costs they have incurred. As in Studies 1a and 1b, we subtracted the direct cost of the resource stolen or damaged by the transgressor from the value of fines (see Fig. 11). Consistent with our preregistration, we estimated the average additional payment for each resource-condition combination using least-squares means. These means were derived from a linear mixed-effects model that predicted the (log-transformed) value of the additional payment as a function of resource type (hogs, cows, vegetables), condition (control, extra-harm, repeated, accidental), and their interaction, with a random intercept for participant ID. The findings conceptually replicate those of Studies 1a and 1b (Fig. 11).

First—consistent with our model—the payments beyond restitution were significantly greater than zero for all resource-condition combinations—hog control: $M = 5.42$, $SE = 0.14$, $t(958.84) = 37.97$, $p < .001$; garden control: $M = 4.20$, $SE = 0.14$, $t(958.84) = 29.43$, $p < .001$; cow control: $M = 5.96$, $SE = 0.143$, $t(958.84) = 41.71$, $p < .001$; all other $ps < .001$ (see Table S16).

Second—contrary to prediction—payments beyond restitution were not equal for all types of resources. Like

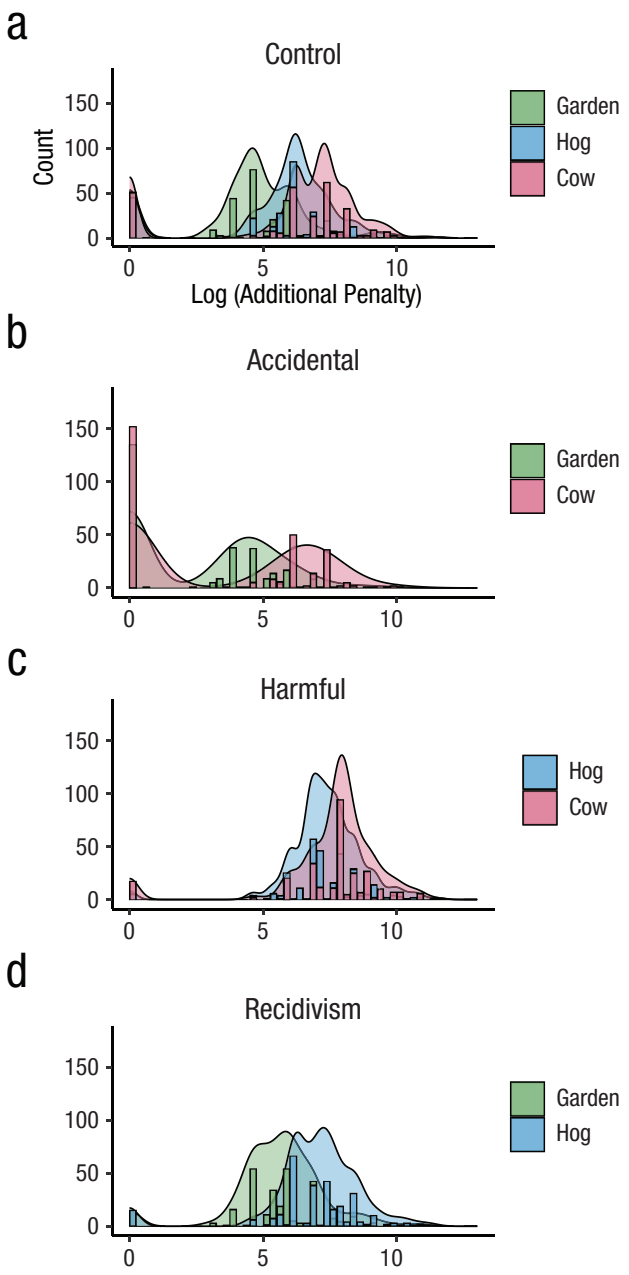

Fig. 11. Distributions of payments prescribed beyond restitution of direct cost of the lost resource in Study 2a (U.S. sample). Control transgressions are illustrated in (a), accidental transgressions in (b), extra-harm transgressions in (c), and repeated transgressions in (d). A value of 0 means that the fine was equal to the value of the victim's loss (e.g., paying \$500 for stealing a pig valued at \$500). Sixty-one fines lower than the value of the resource damaged are also included here as 0.

Mentawai participants in Studies 1a and 1b, instead of applying a constant additional penalty for all kinds of resources, U.S. participants required higher additional penalties for more costly resources (see Fig. 11). In the control conditions, for example, the mean additional payments for stolen or damaged gardens, a resource described as worth \$100, was significantly lower than for damaged cows worth \$1,500 (estimated mean difference: $M_{diff} = -1.75$, $SE = 0.15$, $t(2312) = -11.85$, $p < .001$) and for damaged hogs worth \$500 (estimated mean difference: $M_{diff} = -1.22$, $SE = 0.15$, $t(2312) = -8.24$, $p < .001$). See Table S19 for pairwise comparisons across all resource-condition combinations.

Third—consistent with our model—the additional payment was zero for almost half of the fines in the accidental conditions (44.5%; see Fig. 11). That is, a large proportion of participants considered accidental transgressions to only require repayment of the lost resource and not more. This proportion was much lower for control transgressions (15.7% of the fines), extra-harm transgressions (0.2% of the fines), and recidivist transgressions (4.8% of the fines). An additional 61 fines (2.34% of all fines), most of them also in the accidental conditions (47.5%), were valued as being below the price of strict restitution.

Participants report lower outrage and fines for accidental transgressions and higher outrage and fines for recidivist transgressions. To estimate the effects of recidivism and intentionality on retributive demands, we used the same mixed-effects linear models used above to estimate the effect of extra-harm. As preregistered, these models included condition (control, extra-harm, accidental, repeated) and resource type (hog, cow, garden) as fixed effects, with a random intercept for participant ID. These predictors were used to predict both moral outrage and prescribed fine value (log-transformed).

As shown in Figure 12, accidental transgressions elicited significantly lower outrage, $b = -2.10$, $SE = 0.05$, $t(2291) = -40.29$, $p < .001$, and significantly lower fines, $b = -0.44$, $SE = 0.03$, $t(2315) = -13.33$, $p < .001$, compared with intentional transgressions. As shown in Figure 13, repeat transgressions elicited significantly higher outrage, $b = 0.38$, $SE = 0.05$, $t(2291) = 7.25$, $p < .001$, and significantly higher fines, $b = 0.57$, $SE = 0.03$, $t(2315) = 17.14$, $p < .001$, compared with first-time transgressions. Exploratory analyses revealed that these results are robust to including interaction terms between condition and resource type as additional predictors in both models (see Tables S16 and S17 for full statistics).

Do fines track the predicted reputational cost of transgressions? Our model predicts that, after compensating for direct costs, retributive demands should be highest for recidivist transgressions (because they impose the greatest reputational costs) and lowest for unintentional offenses (because they impose the lowest reputational costs); extra-harm transgressions should be

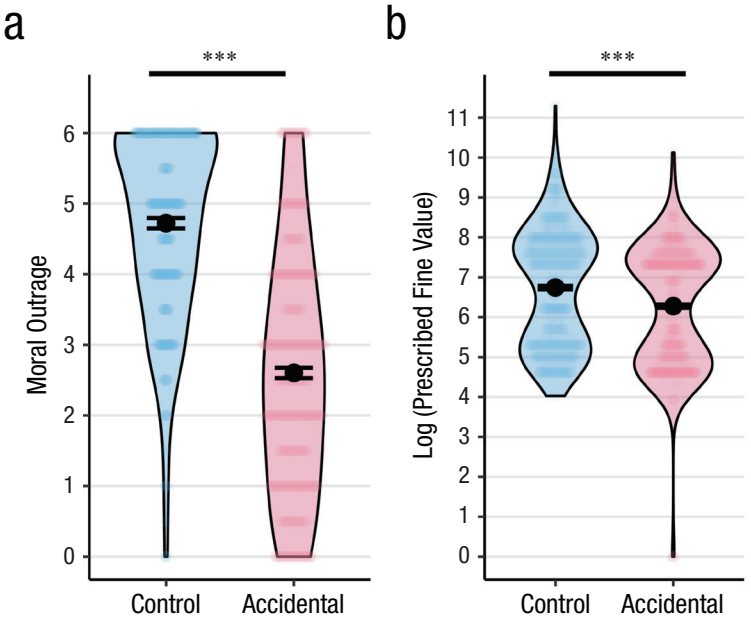

Fig. 12. Differences between the control transgressions and the accidental transgressions for (a) moral outrage ratings and (b) prescribed fine value (log-transformed) in Studies 2a and 2b (U.S. participants). Error bars represent the standard errors of the least-squares means estimated from a linear mixed-effects model with condition and resource as fixed effects and participant as a random intercept. Note that the control conditions here exclude the control transgressions involving hogs because these do not have an accidental equivalent to which they could be compared; there were no accidental transgressions involving hogs in our design (see the Method section). ***$p < .001$.

equivalent to those in the control condition after compensating for the direct cost, as the reputational cost is kept the same across both treatments (Fig. 2).

As preregistered, we regressed the prescribed fine value (log-transformed) on resource type (pigs, chickens, coconut trees), condition (control, extra-harm, unintentional, repeated), and the direct cost of the transgression, with a random intercept for participant ID. After accounting for the compensation of direct costs, participants demanded significantly higher fines for recidivist transgressions than for control transgressions, $b = 0.59$, $SE = 0.03$, $t(2314) = 17.44$, $p < .001$, and significantly lower fines for unintentional transgressions, $b = -0.46$, $SE = 0.03$, $t(2314) = -13.69$, $p < .001$. However, unlike in Studies 1a and 1b, fines were also significantly higher for extra-harm transgressions compared with control transgressions, $b = 0.83$, $SE = 0.06$, $t(2314) = 14.80$, $p < .001$. These results were robust when we included an interaction term between condition and resource type in the model (see Table S18).

General Discussion

The psychology of punishment poses a paradox: Standard evolutionary models suggest that punitive sentiments evolved for deterrence (McCullough et al., 2013), yet psychological evidence finds that people are intuitive retributivists, aiming more to "restore a balance" between offender and victim than to maximize deterrence (Twardawski et al., 2022). Across four experiments, we found evidence consistent with a possible solution, namely that retributive sentiments track the restoration of mutual benefit between cooperative partners—which requires both compensating the victim and deterring offenders (Fig. 1; André et al., 2022; Fitouchi & Singh, 2023). We tested predictions of this idea in a small-scale society without centralized justice institutions (Mentawai horticulturalists, Indonesia) and conceptually replicated the results with a larger sample of U.S. participants.

As predicted by the restorative model, we find that retributive sentiments track both deterrent and compensatory concerns. Consistent with compensatory concerns, Mentawai participants demanded that transgressors pay victims in the currency of their loss, and both Mentawai and U.S. participants were more outraged, and prescribed higher fines, for transgressions that imposed a higher cost on the victim for the same benefit to the transgressor. On the other hand, consistent with deterrence concerns emphasized by evolutionary models (McCullough et al., 2013), both Mentawai and U.S. participants demanded

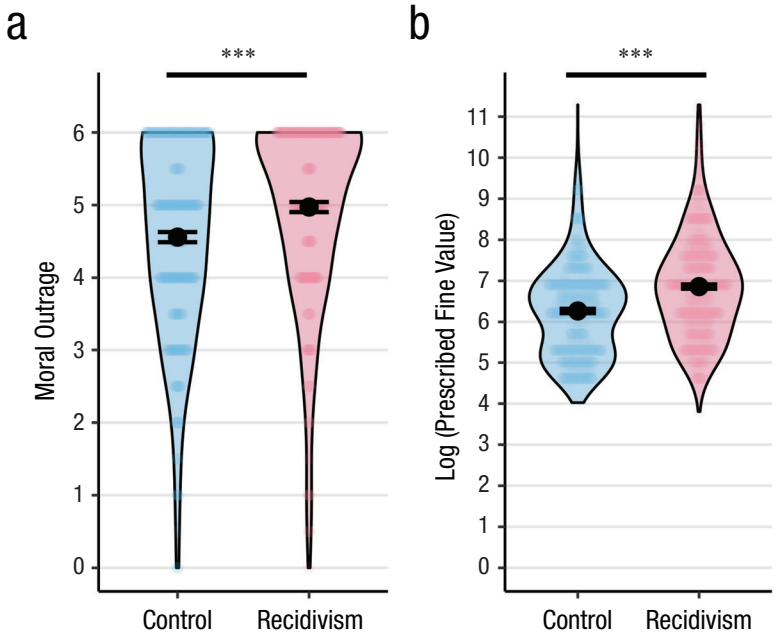

Fig. 13. Differences between the control transgressions and the recidivist transgressions for (a) moral outrage ratings and (b) prescribed fine value (log-transformed) in Studies 2a and 2b (U.S. participants). Error bars represent the standard errors of the least-squares means estimated from a linear mixed-effects model, with condition and resource as fixed effects and participant as a random intercept. Note that the control conditions here exclude the control transgressions involving cows because they did not have a recidivist equivalent to which they could be compared; there were no recidivist transgressions involving cows in our design (see the Method section). ***$p < .001$.

payments that exceeded strict restitution and that were higher for transgressions that inflicted a greater reputational cost (repeat offenses) and lower for transgressions that inflicted a lower reputational cost (accidental offenses). This suggests that deterrence concerns may be more important than previously assumed by the intuitive retributivism literature. In fact, some Mentawai described the additional fine levied on recidivists as *kiroinia*, or "its *kiroi*." *Kiroi* implies learning in this context, explicitly denoting a discouraging function. This fusion of deterrent and compensatory concerns is consistent with previous arguments and ethnographic evidence that restoring relationships is a key challenge after exploitative acts in smaller-scale societies (Fitouchi & Singh, 2023; Petersen et al., 2012; Wiessner, 2020).

Our findings also bear on the role of intentions in punitive judgments. Moral psychologists have long found that WEIRD participants judge unintentional transgressions less harshly than intentional ones (Cushman, 2008). On the basis of cross-cultural experiments (Barrett et al., 2016), some researchers have argued that WEIRD societies are exceptional in this regard and that societies organized around strong kinship networks consider intentions less when making moral judgments (Curtin et al., 2020). We found that Mentawai participants, whose social lives are largely governed by patrilineal clans, do take intentions into account, just like our U.S. participants. Participants in both samples were less outraged by accidental offenses, and many even demanded that the transgressor simply replace the lost resource and pay no additional fine when an act was accidental as opposed to intentional. More broadly, the striking similarities across results between U.S. and Mentawai participants—two societies with profoundly different cultures, justice systems, and subsistence strategies—support the view that central features of people's punitive psychology arise from a shared human nature rather than culturally specific norms (Sznycer & Patrick, 2020).

Some of our findings suggest a need to update the model presented in the introduction. First, we hypothesized that the additional penalty for compensating the reputational cost is a constant added to the direct cost of the transgression (Fig. 1). Yet participants in both samples demanded greater additional penalties for more costly transgressions (Figs. 7 and 11). In fact, a common Mentawai adage—*sabbek silinia, sabbek tulounia* ("one [is] its exchange, one [is] its fine")—dictates that the value of a fine should be equivalent to twice the value of the

destroyed or stolen property. This suggests that transgressions with higher direct cost also impose a higher reputational cost. Second, we hypothesized the additional penalty beyond restitution to be null for accidental transgressions (Fig. 2), yet many participants demanded substantial additional penalties for those transgressions (Figs. 7 and 11). One possibility is that accidents are perceived as causing some reputational cost because of uncertainty over whether they are 100% accidental or instead involve some degree of carelessness or partial intent. An alternative possibility is that people punish accidents to reinforce good behavior regardless of intent (Sarin & Cushman, 2024) or to level fitness differentials regardless of wrongdoing (Deutchman et al., 2021; Price et al., 2002).

This research has several limitations. First, our Mentawai sample size was small given the constraints of acquiring experimental data on participants in this setting. Second, the transgressions we studied were interpersonal violations, sometimes referred to as "torts" or "private grievances," as opposed to violations that harm the welfare of society as a whole, sometimes referred to as "crimes" (Solomon, 2010). People in various societies, including some small-scale communities, seem to distinguish between these two kinds of transgressions and to punish them differently (Fitouchi & Singh, 2023; Richardson, 1940), limiting our ability to generalize the features of punishment documented in this context to the punishment of all transgressions. Third, although previous research documents the importance of power asymmetries in shaping retributive sentiments (Sell et al., 2009; Von Rueden & Gurven, 2012), our design controlled for such variation and therefore did not investigate this important factor. Finally, this research analyzed retributive sentiments in only two populations. Although our participants come from societies with drastically different justice systems, human societies exhibit a profound diversity of justice institutions, plausibly reflecting (or contributing to) an equivalent diversity of retributive sentiments. Regardless, our research demonstrates the value of studying punitive intuitions outside of the WEIRD experimental settings in which they are normally assessed.

Transparency

Action Editor: Yoel Inbar
Editor: Simine Vazire
Author Contributions

Léo Fitouchi: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Visualization; Writing – original draft; Writing – review & editing.

Manvir Singh: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Visualization; Writing – original draft; Writing – review & editing.

Léo Fitouchi and Manvir Singh contributed equally.

Declaration of Conflicting Interests

The authors declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

L. Fitouchi acknowledges funding from the French Agence Nationale de la Recherche (under the Investissement d'Avenir programme, ANR-17-EURE-0010). Field research was supported by a Graduate Research Fellowship from the National Science Foundation, a Sheldon Traveling Fellowship from the Harvard Committee on General Scholarships, and a grant from the Mind, Brain, and Behavior Initiative at Harvard University.

Artificial Intelligence

Artificial-intelligence-assisted technologies were used only to rephrase sentences and revise some code in this research. No other artificial-intelligence-assisted technologies were used in this research or the creation of this article.

Ethics

Studies 1a and 1b were approved by the Harvard University Committee on the Use of Human Subjects (ID No. IRB16-1080) and the Ministry of Research, Technology and Higher Education of the Republic of Indonesia. Studies 2a and 2b were approved by the Review Board for Ethical Standards in Research of the Toulouse School of Economics and Institute for Advanced Study in Toulouse.

Open Practices

Open practices for this article are described in the Research Transparency Statement section, which appears at the end of the Introduction section in the main text.

ORCID iDs

Léo Fitouchi https://orcid.org/0000-0002-1724-0129
Manvir Singh https://orcid.org/0000-0002-6026-947X

Acknowledgments

We thank B. Sakaliou, L. Glowacki, M. Grote, and S. Worthington at the Harvard Institute for Quantitative Social Science for their assistance at various stages of this project. F. Cushman and members of the Evolution and Ecology of Human Behavior and Culture Group at the University of California, Davis provided valuable comments on an earlier draft of this manuscript.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976261439088

References

André, J.-B., Fitouchi, L., Debove, S., & Baumard, N. (2022). *An evolutionary contractualist theory of morality*. OSF. https://doi.org/10.31234/osf.io/2hxgu

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M. T., Fitzpatrick, S., Gurven, M., Henrich, J., Kanovsky, M.,

Kushnick, G., Pisor, A., Scelza, B. A., Stich, S., Von Rueden, C., Zhao, W., & Laurence, S. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences*, *113*(17), 4688–4693. https://doi.org/10.1073/pnas.1522070113

Baumard, N. (2012). The restorative logic of punishment: Another argument in favor of weak selection. *Behavioral and Brain Sciences*, *35*(1), 17–18.

Black, D. J. (1993). *The social structure of right and wrong*. Academic Press.

Boyd, R. (2018). *A different kind of animal: How culture transformed our species*. Princeton University Press.

Carlsmith, K. M. (2008). On justifying punishment: The discrepancy between words and actions. *Social Justice Research*, *21*(2), 119–137. https://doi.org/10.1007/s11211-008-0068-x

Carlsmith, K. M., & Darley, J. M. (2008). Psychological aspects of retributive justice. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 40, pp. 193–236). Elsevier. https://doi.org/10.1016/S0065-2601(07)00004-4

Curtin, C. M., Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M. T., Fitzpatrick, S., Gurven, M., Kanovsky, M., Kushnick, G., Laurence, S., Pisor, A., Scelza, B., Stich, S., Von Rueden, C., & Henrich, J. (2020). Kinship intensity and the use of mental states in moral judgment across societies. *Evolution and Human Behavior*, *41*(5), 415–429. https://doi.org/10.1016/j.evolhumbehav.2020.07.002

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006

Deutchman, P., Bračič, M., Raihani, N., & McAuliffe, K. (2021). Punishment is strongly motivated by revenge and weakly motivated by inequity aversion. *Evolution and Human Behavior*, *42*(1), 12–20. https://doi.org/10.1016/j.evolhumbehav.2020.06.001

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

Fessler, D. M. T., Barrett, H. C., Kanovsky, M., Stich, S., Holbrook, C., Henrich, J., Bolyanatz, A. H., Gervais, M. M., Gurven, M., Kushnick, G., Pisor, A. C., von Rueden, C., & Laurence, S. (2015). Moral parochialism and contextual contingency across seven societies. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1813), Article 20150907. https://doi.org/10.1098/rspb.2015.0907

Fitouchi, L., André, J.-B., & Baumard, N. (2024). Are there really so many moral emotions? Carving morality at its functional joints. In L. Al-Shawaf & T. K. Shackelford (Eds.), *The Oxford handbook of evolution and the emotions* (p. 944–967). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197544754.013.35

Fitouchi, L., & Singh, M. (2023). Punitive justice serves to restore reciprocal cooperation in three small-scale societies. *Evolution and Human Behavior*, *44*(5), 502–514. https://doi.org/10.1016/j.evolhumbehav.2023.03.001

Gardner, A., & West, S. A. (2004). Cooperation and punishment, especially in humans. *The American Naturalist*, *164*(6), 753–764. https://doi.org/10.1086/425623

Krams, I., Kokko, H., Vrublevska, J., Āboliņš-Ābols, M., Krama, T., & Rantala, M. J. (2013). The excuse principle can maintain cooperation through forgivable defection in the Prisoner's Dilemma game. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1766), Article 20131475. https://doi.org/10.1098/rspb.2013.1475

Levine, S., Chater, N., Tenenbaum, J. B., & Cushman, F. (2024). Resource-rational contractualism: A triple theory of moral cognition. *Behavioral and Brain Sciences*, 1–38. https://doi.org/10.1017/S0140525X24001067

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences*, *36*(1), 1–15. https://doi.org/10.1017/S0140525X11002160

Miller, W. I. (2005). *Eye for an eye*. Cambridge University Press. https://books.google.fr/books?hl=fr&lr=&id=_RMiOXoLnncC&oi=fnd&pg=PR9&dq=eye+for+an+eye+miller&ots=hafacvouzA&sig=7gLyD5qXd_odOe7VEuaMv-vfvxQ

Molho, C., Twardawski, M., & Fan, L. (2022). What motivates direct and indirect punishment?: Extending the "intuitive retributivism" hypothesis. *Zeitschrift Für Psychologie*, *230*(2), 84–93. https://doi.org/10.1027/2151-2604/a000455

Persoon, G. (1987). Two civil servants leaving the Island of Siberut (Mentawi Archipelago, Indonesia). In R. de Ridder, & J. A. J. Karremans (Eds.), *The Leiden tradition in structural anthropology: Essays in honour of P. E. de Josselin de Jong* (p. 140). E. J. Brill, Leiden.

Petersen, M. B., Sell, A., Tooby, J., & Cosmides, L. (2012). To punish or repair? Evolutionary psychology and lay intuitions about modern criminal justice. *Evolution and Human Behavior*, *33*(6), 682–695. https://doi.org/10.1016/j.evolhumbehav.2012.05.003

Price, M. E., Cosmides, L., & Tooby, J. (2002). Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior*, *23*(3), 203–231. https://doi.org/10.1016/S1090-5138(01)00093-9

Richardson, J. (1940). *Law and status among the Kiowa Indians* (*Vol. 1*). University of Washington Press.

Sarin, A., & Cushman, F. (2024). Once thought too few: An adaptive rationale for punishing negligence. *Psychological Review*, *131*(3), 812.

Schefold, R. (1988). *Lia: Das grosse Ritual auf den Mentawai-Inseln (Indonesien)*. D. Reimer. https://ehrafworldcultures.yale.edu/document?id=od09-010

Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences*, *106*(35), 15073–15078. https://doi.org/10.1073/pnas.0904312106

Singh, M., & Garfield, Z. H. (2022). Evidence for third-party mediation but not punishment in Mentawai justice. *Nature Human Behaviour*, *6*(7), 930–940. https://doi.org/10.1038/s41562-022-01341-7

Singh, M., Kaptchuk, T. J., & Henrich, J. (2021). Small gods, rituals, and cooperation: The Mentawai water spirit Sikameinan. *Evolution and Human Behavior*, *42*(1), 61–72. https://doi.org/10.1016/j.evolhumbehav.2020.07.008

Solomon, J. M. (2010). What is civil justice. *Loyola of Los Angeles Law Review*, *44*, 317–337.

Sznycer, D., & Patrick, C. (2020). The origins of criminal law. *Nature Human Behaviour*, *4*(5), 506–516. https://doi.org/10.1038/s41562-020-0827-8

Sznycer, D., Sell, A., & Williams, K. E. G. (2023). Justice-making institutions and the ancestral logic of conflict. *Evolution and Human Behavior*, *44*(3), 184–194. https://doi.org/10.1016/j.evolhumbehav.2022.12.003

Tulius, J. (2012). *Family stories: Oral tradition, memories of the past, and contemporary conflicts over land in Metawai- Indonesia* [Leiden University]. https://hdl.handle.net/1887/20262

Twardawski, M., Gollwitzer, M., Pohl, S., & Bošnjak, M. (2022). What drives second- and third-party punishment? *Zeitschrift Für Psychologie*, *230*(2), 77–83. https://doi.org/10.1027/2151-2604/a000454

Von Rueden, C. R., & Gurven, M. (2012). When the strong punish: Why net costs of punishment are often negligible. *Behavioral and Brain Sciences*, *35*(1), 43–44. https://doi.org/10.1017/S0140525X11001427

Wiessner, P. (2020). The role of third parties in norm enforcement in customary courts among the Enga of Papua New Guinea. *Proceedings of the National Academy of Sciences*, *117*(51), 32320–32328. https://doi.org/10.1073/pnas.2014759117

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*(20), 8235–8240. https://doi.org/10.1073/pnas.0701408104