# The Value of vengeance and the demand for deterrence

Crockett, Molly J ; Ozdemir, Yagiz ; Fehr, Ernst

Abstract: Humans will incur costs to punish others who violate social norms. Theories of justice highlight 2 motives for punishment: a forward-looking deterrence of future norm violations and a backward-looking retributive desire to harm. Previous studies of costly punishment have not isolated how much people are willing to pay for retribution alone, because typically punishment both inflicts damage (satisfying the retributive motive) and communicates a norm violation (satisfying the deterrence motive). Here, we isolated retributive motives by examining how much people will invest in punishment when the punished individual will never learn about the punishment. Such "hidden" punishment cannot deter future norm violations but was nevertheless frequently used by both 2nd-party victims and 3rd-party observers of norm violations, indicating that retributive motives drive punishment decisions independently from deterrence goals. While self-reports of deterrence motives correlated with deterrence-related punishment behavior, self-reports of retributive motives did not correlate with retributive punishment behavior. Our findings reveal a preference for pure retribution that can lead to punishment without any social benefits. (PsycINFO Database Record (c) 2014 APA, all rights reserved).

DOI: https://doi.org/10.1037/xge0000018

# The Value of Vengeance and the Demand for Deterrence

Molly J. Crockett[1,*], Yagiz Özdemir[1,*], & Ernst Fehr[1]

[1]Laboratory for Social and Neural Systems Research, Department of Economics, University of Zurich

[*]equal contribution

## Abstract

Humans will incur costs to punish others who violate social norms. Theories of justice highlight two motives for punishment: a forward-looking deterrence of future norm violations, and a backward-looking retributive desire to harm. Previous studies of costly punishment have not isolated how much people are willing to pay for retribution alone, because typically punishment both inflicts damage (satisfying the retributive motive) and communicates a norm violation (satisfying the deterrence motive). Here, we isolated retributive motives by examining how much people will invest in punishment when the punished individual will never learn about the punishment. Such 'hidden' punishment cannot deter future norm violations, but was nevertheless frequently used by both second-party victims and third-party observers of norm violations, indicating that retributive motives drive punishment decisions independently from deterrence goals. While self-reports of deterrence motives correlated with deterrence-related punishment behavior, self-reports of retributive motives did not correlate with retributive punishment behavior. Our findings reveal a preference for pure retribution that can lead to punishment without any social benefits.

Punishment of social norm violations is widespread across human societies (Henrich et al., 2006). Under certain conditions, punishment can prevent free-riding and promote cooperation, and many people are willing to 'altruistically' punish anonymous strangers, even when it is costly and yields no material or reputational benefits (Fehr & Fischbacher, 2003). Yet the motivational basis of costly altruistic punishment is not fully understood. Theories of justice highlight two major proximate motives for punishment: *deterrence* and *retribution* (Bentham, 1970; Kant, 1965). People motivated by deterrence employ punishment to prevent norm violators from repeating their bad behavior in the future; the goal of punishment is to teach a lesson by communicating that a norm has been violated. In contrast, people motivated by retribution employ punishment to cause norm violators to suffer; the goal of punishment is to inflict damage. Although these motives are separate in principle, they are intertwined in practice: any punishment that is communicated to the punisher satisfies both deterrence and retribution goals because it communicates a norm violation and the existence of people who are willing to punish (both of which may reduce future norm violations); and it inflicts damage to the norm violator (satisfying the retributive goal).

Understanding the extent to which punishment is driven by retributive motives has potentially important implications for the design of public institutions to promote social norms. If individuals derive private satisfaction from punishment irrespective of its ability to deter future harms, they may

utilize punishment inefficiently in terms of promoting social welfare by, for instance, persisting in punishment even in cases where its future benefits are limited.

Previous studies of punishment motives are consistent with the view that people are concerned about both deterrence and retribution. When asked to provide justifications for punishment, people frequently report a motivation to deter future crimes (Ellsworth & Ross, 1983; Vidmar & Miller, 1979). In hypothetical scenarios, punishment decisions are more sensitive to factors that are primarily associated with retribution (e.g., the severity of the crime) than to factors associated with deterrence (e.g., the likelihood of future transgressions) (Carlsmith, 2006; Carlsmith, Darley, & Robinson, 2002). This work provides evidence that both retribution and deterrence motives may play a role in punishment decisions, but based on these studies it remains unclear to what extent people are willing to invest their own resources in punishment that fulfills retribution vs. deterrence goals.

More recent studies of costly punishment have demonstrated that people are indeed willing to sacrifice personal payoffs in order to reduce the payoffs of norm violators (Fehr & Fischbacher, 2003). However, these studies have not disentangled the communication of norms and the infliction of damage. It therefore remains unknown to what extent humans will invest their own resources to deter future norm violations versus to exact retribution. In other words, behavioral evidence for costly pure retribution in humans is lacking: it is not known whether individuals are willing to bear the cost of purely retributive sanctions. There is some evidence hinting that people may

be willing to pay for retribution alone; punishment levels are substantial even when the implementation of punishment is delayed until after all interaction is over (Fudenberg & Pathak, 2010), and in one-shot games when there is no opportunity for future interactions in the laboratory (Fehr & Fischbacher, 2003). However, the potential effects of punishment on future behavior may well extend beyond the specific context of the laboratory: subjects who are informed that they are punished for a norm violation in a lab experiment may reduce future norm violations in similar situations outside the lab. Finally, neuroimaging studies have demonstrated activity in reward circuitry, including the striatum and medial prefrontal cortex (mPFC), during punishment of norm violators (Crockett et al., 2013; de Quervain et al., 2004), consistent with the notion that humans derive pleasure from punishment. But since the striatum and mPFC are known to be involved in anticipating distant rewarding outcomes (Kable & Glimcher, 2007), as well as encoding immediately rewarding outcomes (Haber & Knutson, 2009), these studies cannot rule out the possibility that punishment-related responses in these regions reflect the expected social benefits of deterring future norm violations. Moreover, striatal responses during punishment do not necessarily indicate feelings of pleasure (Poldrack, 2006), as the striatum is sometimes also involved in processing aversive outcomes (Delgado, Li, Schiller, & Phelps, 2008).

An additional question concerns differences in punishment motives between second parties who are affected by the norm violation and unaffected third parties. Empirical evidence shows that second parties punish more strongly than unaffected third parties (Fehr & Fischbacher, 2004), and

the prevalence of third-party norm enforcement institutions such as juries concords with the common notion that third parties ought to punish in a more impartial or normative manner (Hirsch, 1986; Tunick, 1992). However, the extent to which retributive motives differ between second- and third-party punishment remains unclear. Comparing second- and third-party punishment is not straightforward, however. Previous attempts suffer from an obvious confound: in second-party punishment only two players are involved, whereas in third-party punishment three players are involved. This is potentially problematic because punishment decisions are sensitive to the presence of an audience (Kurzban, DeScioli, & O'Brien, 2007; Piazza & Bering, 2008). We addressed this issue by examining both second- and third- party punishment in a three-player setting. Our goal in the current study was to characterize the extent of proximate motives for retribution and deterrence in second- and third-party punishment.

**Participants**

Two hundred and fifty-nine healthy volunteers provided informed consent and participated in the study that was approved by the ethics committee of the Department of Economics, University of Zürich. One hundred and eleven healthy male volunteers (mean age: 23.2 y) participated in the role of player P, whose behavior was the main focus of the current study. These participants attended testing sessions in the Economics Laboratory at the University of Zürich, for which they received a participation fee of CHF 25, plus an additional payment based on their decisions in the study.

## Method

### *Three-player trust game with punishment*

In our basic setting, three players (a punisher, P; a bystander, B; and a trustee, T) interact anonymously with each other. The punisher (P) and the bystander (B) each receive an endowment of CHF 5. The game has three stages. In the *trust* stage, P and B can entrust their endowment to T. Each entrusted endowment is multiplied by a multiplier *m* and transferred to T. Trustees were instructed that the multiplier could be any integer value between 2 and 6.

In the *back-transfer* stage, T decides what proportion (0%, 25%, or 50%) of the received endowment (CHF 5 * m) to send back to *one* of the players, either P (second-party punishment, Fig. 1a) or B (third-party punishment, Fig. 1b). For the remaining player, the computer decides T's back-transfer. Thus, in the second-party punishment condition, T decides how to repay P's trust, while the computer determines how T repays B's trust. In the third-party punishment condition, T decides how to repay B's trust, while the computer determines how T repays P's trust.

Finally in the *punishment* stage, P receives an additional endowment of CHF 5 and is able to spend up to his entire endowment to reduce T's payoff; each CHF 0.10 spent by P resulted in a payoff reduction of CHF 0.20 for T (see SOM for details).

In sequential trust and social dilemma games, a strong norm of conditional cooperation applies (Fehr & Fischbacher, 2003; 2004). This norm

demands that T responds kindly to initial cooperative acts of P and B in the first stage. Intentional back-transfers of 0% in the second stage therefore unambiguously violate this norm. In the second party condition only P is the victim of such a norm violation, while in the third party condition only B is the victim. We therefore expected P to punish T for intentional back-transfers of 0%, since these represent norm violations.

***Isolating retributive motives***

We isolated retributive motives by tightly controlling T's knowledge of whether he has been punished across two key experimental conditions. Although T's payoff is always reduced when P punishes him (and P knows this), whether T *learns* about the punishment varies across conditions. In the *open* punishment condition, T receives a written message informing him whether P has punished him. In the *hidden* punishment condition, T is *not* informed whether P has punished him. This was made explicit to the P players in the experimental instructions, and P players had to pass a comprehension quiz to demonstrate their understanding of this before they started the decision-making phase of the experiment.

We were able to control T's knowledge about his punishment in several ways. T was not informed of the size of the total endowment that he received through P's and B's transfers, nor the size of the back-transfer determined by the random device. Moreover, because a specific final payoff in the technically possible payoff range could arise in many different ways, the final payoff also

provided no information about punishment (see SI for a detailed explanation).

We used detailed instructions to ensure that the punisher P was aware of the difference between open and hidden punishment when he made his punishment decisions. We confirmed this with a comprehension quiz (see SOM for details).

Our experimental design provides a stringent test for the existence of retributive motives in humans. The hidden punishment condition excludes the deterrence motive, because deterring future norm violations requires that the perpetrator knows that he has been punished. Thus, higher punishment of unfair back-transfers (relative to fair back-transfers) in the hidden condition reflects retributive motives, i.e., the private satisfaction derived from reducing the payoff of a norm violator. In contrast, higher punishment of unfair back-transfers (relative to fair back-transfers) in the open condition reflects a combination of retribution and deterrence motives (Table 1). Because the open condition has the same retributive effects as the hidden condition, but with the added benefit of deterrence, we expected open punishment of unfair back-transfers to be both more likely and more substantial than hidden punishment of unfair back-transfers. And based on previous studies suggesting a potential role of retributive motives in punishment (Carlsmith, 2006; Carlsmith et al., 2002), we expected to observe higher punishment of unfair back-transfers (relative to fair back-transfers) in the hidden condition, despite the fact that unambiguous behavioral evidence for pure retribution is currently lacking.

### *Controlling for payoff-based motives*

Decisions to punish can also be motivated by inequality aversion (Fehr & Schmidt, 1999) or other types of payoff-based social preferences such as spite (Jensen, 2010). People who dislike inequality will punish others with higher payoffs, regardless of whether the target of punishment is not responsible for payoff allocations (Blount, 1995; Falk, Fehr, & Fischbacher, 2008). Likewise, spiteful subjects punish regardless of whether the trustee decided intentionally or whether a random device determined the back-transfer (Table 1). To separately control for such motives, we implemented a "computer control" condition in which T's back-transfer decisions vis à vis both P and B were unintentional (i.e., determined by the computer; Fig. 1c). In the computer control condition, punishers faced a set of decisions that were identical to the two experimental conditions in all respects aside from the intentionality of the trustee T (Fig. S1, SOM).

### *General procedure*

We collected the decisions of B and T players in advance (see SOM), so that we were able to face each player P with an identical set of games without using deception. Each punisher P played a series of 54 anonymous one-shot trust games with punishment, each with different individuals in the roles of B and T. Each player P faced the same set of 54 games, reflecting a factorial within-subjects design that crossed (a) level of T's back-transfer (0%, 25%, or 50%), (b) second- vs. third-party punishment, (c) whether punishment was open or hidden, and (d) whether T's back-transfer was intentional or

unintentional (Figure S1). The dependent measure was the amount P spent on punishment in each game. Subjects had unlimited time to make their punishment decisions. Punishment decision data were analyzed in SPSS 18 using the generalized estimating equations (GEE) procedure, which generates for each tested main effect and interaction a chi-square statistic, a 95% confidence interval, and an associated p-value. We used an independent working correlation matrix given that participants played one-shot games and thus the correlation between repeated measurements should be low. For analysis of binary (yes/no) punishment decisions, we used a logistic link function, and for analysis of continuous punishment amounts we used a linear link function. Effect sizes were computed using Cohen's *d*.

Following the 54 games, participants completed a questionnaire concerning their motivations for punishment (see SI). Both the games and the questionnaire were implemented using z-Tree (Fischbacher, 2007). At the end of the session, one of the 54 games was randomly selected for payment for each subject. Subjects in the role of P received their payments in cash immediately. Subjects in the roles of T and B whose decisions were implemented in the randomly selected game received their payments by post. If the randomly selected game was one with open punishment, the payment sent to T included a letter that revealed whether P punished T, and by how much.

## Results

### *Retribution and deterrence in second-party punishment*

In second-party punishment trials, P decided whether and how much to punish T for intentionally sending back 0%, 25%, or 50% of the money to P. As expected, back-transfer level had a significant effect on second-party punishment (likelihood: $\chi^2$=16.781, p<0.001, d = 0.84; amount: $\chi^2$=19.663, p<0.001, d = 0.93); P was much more likely to punish, and spent more to punish, T when he sent back 0% of the money, relative to 25% and 50%. Critically, subjects distinguished between fair and unfair back-transfers in both the open condition (likelihood: $\chi^2$=24.907, p<0.001, d = 1.08; amount: $\chi^2$=21.673, p<0.001, d = 0.99; Fig. 2, striped red bars) and the hidden condition (likelihood: $\chi^2$=9.544, p=0.008, d = 0.61; amount: $\chi^2$=13.419, p=0.001, d = 0.74; Fig. 2, solid red bars). The latter result provides unambiguous evidence for second-party retributive motives in humans. Finally, in line with our predictions, open punishment was both more likely and more substantial than hidden punishment, particularly for 0% back-transfers (open*back-transfer interaction, likelihood: $\chi^2$=12.487, p=0.002, d = 0.71; amount: $\chi^2$=11.419, p=0.003, d = 0.68). These findings demonstrate that the preference to communicate norms through punishment also plays an important role for punishment decisions.

***Retribution and deterrence in third-party punishment***

In third-party punishment trials, P decided whether and how much to punish T for intentionally sending back 0%, 25%, or 50% of the money entrusted to him by B. We found that participants were less likely to engage in third-party punishment than second-party punishment ($\chi^2$=15.501, p<0.001, d

= 0.81), and spent less on third-party punishment than second-party punishment ($\chi^2$=10.505, p=0.001, d = 0.65). Thus, third party punishment is less likely and less strong even when controlling for the number of players involved in the interaction.

To what extent did retribution motivate third-party punishment? Similar to second-party punishment, we observed a main effect of T's back-transfer to B on P's decisions to punish T (likelihood: $\chi^2$=16.049, p<0.001, d = 0.82; amount spent: $\chi^2$=12.856, p<0.001, d = 0.72). Again, subjects distinguished between fair and unfair back-transfers in both the open condition (likelihood: $\chi^2$=18.266, p<0.001, d = 0.89; amount: $\chi^2$=12.019, p=0.002, d = 0.70; Fig. 2, striped blue bars) and the hidden condition (likelihood: $\chi^2$=8.122, p=0.017, d = 0.56; amount: $\chi^2$=5.909, p=0.052, d = 0.47; Fig. 2, solid blue bars), providing evidence for third-party retributive motives. Finally, as was the case for second-party punishment, open punishment was both more likely and more substantial than hidden punishment, across all levels of back-transfer (main effect of open, likelihood: $\chi^2$=5.542, p=0.019, d = 0.46; amount: $\chi^2$=10.915, p=0.002, d = 0.66). The effect of norm communication on punishment of unfair back-transfers was no larger for third-party punishment than for second-party punishment (party*open*back-transfer interaction, likelihood: $\chi^2$=0.613, p=0.736; amount: $\chi^2$=2.211, p=0.331).

***Controlling for payoff-based motives***

One potential alternative explanation for the observation of hidden punishment is that such punishment reflects inequality aversion, spite or other

types of purely payoff-based social preferences rather than retributive motives. Note that retributive motives can only play a role when back-transfers are intentional while the punisher's payoff-based social preferences might play a role in the punishment of both intentional and unintentional back-transfers. Therefore, we can rule out these alternative explanations by comparing hidden punishment of intentional back-transfers by T with hidden punishment of unintentional back-transfers by T (matched for amount). In computer control trials (Fig. 1c), the computer decided player T's back-transfers to both P and B; therefore, in these trials, player T was not responsible for the level of back-transfer. Thus, the observation of higher punishment in the hidden-intentional condition, relative to the hidden-unintentional condition, constitutes evidence for retributive motives over and above purely payoff-based social preferences.

We observed significantly more punishment in the hidden-intentional condition, relative to the hidden-unintentional condition. For second-party hidden punishment, there was a significant main effect of intentionality on punishment (likelihood: $\chi^2$=9.875, p=0.002, d = 0.62; amount: $\chi^2$=10.125, p=0.001, d = 0.63; Fig. 3, red bars); intentional back-transfers were punished more strongly than unintentional ones of equal value. This effect of intentionality was strongest for 0% back-transfers, as evidenced by a significant interaction between intentionality and back-transfer (likelihood: $\chi^2$=7.217, p=0.027, d = 0.53; amount: $\chi^2$=9.525, p=0.009, d = 0.61). For third-party hidden punishment, there was also a significant interaction between intentionality and back-transfer; intentional back-transfers were punished

more strongly than unintentional back-transfers, but only for the most unfair (0%) back-transfers (likelihood: $\chi^2$=6.950, p=0.031, d = 0.52; amount: $\chi^2$=6.732, p=0.035, d = 0.51; Fig. 3, blue bars). Thus, payoff-based motives could not completely explain hidden punishment in either second-or third-party punishment.

We next examined differences in retributive motives between second- and third-party punishment, focusing exclusively on trials in the hidden condition. The average level of hidden punishment of *unintentional* 0% transfers did not differ significantly between second- and third-party conditions ($\chi^2$=1.736, p=0.188), suggesting that second- and third-party punishment were matched in terms of purely payoff-based social preferences. However, the average amount of hidden punishment of *intentional* 0% transfers was significantly greater in second- than third-party punishment ($\chi^2$=7.125, p=0.008, d = 0.52). This observation was confirmed by a significant two-way interaction between party and intentionality ($\chi^2$=4.558, p=0.033, d = 0.41) within the hidden condition; punishment in the hidden-intentional condition, *relative to* the hidden-unintentional condition, was greater in second-party than in third-party punishment. These results suggest that retributive motives, while present in both second- and third-party punishment, are stronger in the former than in the latter.

***Self-reported motives for retribution and deterrence***

We next explored the correspondence between subjects' self-reported motives for punishment and their actual punishment behavior. After they had

made all their decisions, we asked subjects to indicate on a Likert scale the extent to which their punishment decisions were motivated by factors associated with deterrence, and factors associated with retribution (see SI for details of the factor analysis). Endorsement of retributive motives was low, with a mean rating of 1.75 (s.e.=0.13) on a 5-point scale. Endorsement of deterrence motives was significantly higher (mean=3.06, s.e.=0.20, $t_{(110)}$=6.769, p<0.001, d=1.29). We then correlated subjects' self-reported ratings against their own behavior. Our behavioral measure of deterrence motives – the difference between amount spent on open relative to hidden punishment of unfair (0%) back-transfers – was positively correlated with self-reported deterrence motives (r=0.417, p=0.004, d = 0.92). However, our behavioral measure of retributive motives—the amount spent on hidden punishment of intentional relative to unintentional unfair (0%) back-transfers—was not significantly correlated with self-reported retributive motives (r=0.017, p=0.913). The relationship between self-report and behavior was stronger for deterrence motives than for retributive motives (Z=1.96, p=0.05, d = 0.38). In fact, self-reported retributive motives did not significantly predict any aspect of punishment behavior (all p>0.687).

**Discussion**

Our findings provide unambiguous behavioral evidence that people are willing to invest personal resources in pure retribution without the possibility of deterrence. We observed higher punishment of unfair back-transfers than fair back-transfers even in our hidden treatment where the norm-enforcing

properties of punishment were completely removed. Retributive punishment was evident in both second- and third-party punishment settings, and could not be completely explained by inequality aversion or other purely payoff-based preferences such as spite. These results indicate that people value reducing the payoffs of norm violators, even in the absence of any potential future social benefits of punishment.

At the same time, our data suggest that punishers derive additional value from the opportunity to communicate norms. Costly punishment was both more likely and more substantial when the target of punishment would learn that he was punished, controlling for material damage. This finding is consistent with previous work showing that the opportunity to communicate norms (sometimes called 'emotion expression') can serve as a substitute for inflicting material damage (Xiao & Houser, 2005; Yamagishi et al., 2009).

Alternatively, it is possible that the communication of norms is driven to some extent by a retributive desire to inflict emotional damage (in addition to material damage). Some evidence suggests this is indeed the case. Dictators who anticipate receiving a written message from their recipient give significantly higher amounts than those who will not receive a message, indicating that non-material sanctions carry emotional weight (Ellingsen & Johannesson, 2008; Xiao & Houser, 2009). It is therefore possible that the present study underestimated the extent to which retributive motives drive costly punishment.

We provide a novel method for directly comparing second- and third-party punishment within a single setting. Holding constant the number of

players involved in the interaction, the payoff of the punisher, and the relative payoffs between the punisher and the other players, we observed stronger second-party punishment than third-party punishment. Preferences for the communication of norms did not significantly differ between second- and third-party punishment. However, retributive motives were stronger in second- than third-party punishment. This suggests that personal suffering amplifies the demand for retribution, but not the communication of norms.

Notably, subjects' distinction between open and hidden punishment was strongest for the unfair back-transfers. We observed a few instances of 'antisocial' punishment of fair 50% back-transfers (Herrmann et al., 2008; Gächter et al., 2009; Rand et al., 2010; Rand & Nowak, 2011); unlike punishment of unfair back-transfers, the amount of antisocial punishment did not differ between open and hidden conditions. This suggests that antisocial punishment is driven by a desire to inflict damage on fair players, rather than a desire to communicate a norm of non-cooperation. This hypothesis could be tested further using similar methods as in the present study, but in populations with higher occurrences of antisocial punishment (Herrman et al., 2008).

Our methods also enabled us to disentangle punishment motives within subjects. Previous research on costly punishment behavior has not explicitly separated preferences about material payoffs from preferences about the communication of norm violations, since the target of punishment was always informed that he has been punished. Here we were able to measure the relative contributions of both types of preferences to punishment behavior, and to compare behavioral preferences with self-reported motives. Such

comparisons can be valuable because people may lack insight into their own motives (Nisbett & Wilson, 1977), or be reluctant to disclose motivations that are less socially desirable. Consistent with this view, in our study subjects rarely endorsed retributive motives in the self-report questionnaire. Meanwhile, subjects were more likely to endorse motives for deterrence, perhaps because such motives are more socially desirable. Self-reported motives for deterrence were significantly correlated with our behavioral measure of deterrence, but self-reported motives for retribution were not correlated with our behavioral measure of retribution, or indeed any aspect of punishment behavior. Further research is needed to understand the factors that moderate the correspondence between self-reported motives and behavior.

An intriguing open question is whether preferences for retribution versus deterrence depend on distinct neural systems. Punishment decisions engage brain regions involved in the computation of value, including the striatum and mPFC (Baumgartner et al., 2011; Crockett et al., 2013; de Quervain et al., 2004), but also regions involved in forward planning and goal-directed decision-making, including the dorsolateral PFC (Baumgartner et al., 2011; Buckholtz et al., 2008; Sanfey et al., 2003). While activity in the striatum tracks the amount of material damage inflicted by punishment (de Quervain et al., 2004), prefrontal regions may be sensitive to whether punishment is likely to deter future harms (Buckholtz et al., 2008, 2012). Environmental factors such as stress are known to disrupt prefrontal function (Robbins & Arnsten, 2009), and may therefore alter the nature of punishment decisions.

Understanding the influence of the environment on punishment decisions has important implications for the criminal justice system (Danziger, Levav, & Avnaim-Pesso, 2011).

The National Council on Crime and Delinquency has declared that "sentencing should not be based on revenge and retribution" (Hirsch, 1986; Tunick, 1992). This view is consistent with our finding that retributive motives were less forceful in third-party punishment, relative to second-party punishment. However, our findings also cast some doubt on the notion that 'impartial observers' are capable of meting out punishments in a normative manner immune to emotional influences; retributive motives still explained a substantial portion of third-party punishment. This is perhaps not so surprising in light of humans' remarkable capacity for empathy. Observing harm to another engages similar brain regions as those that signal harm to the self (Singer et al., 2004). Thus, if the desire for retribution arises in response to self-directed harm, it may be similarly triggered by harms against others, to the extent that harms against others feel aversive (Batson, Kennedy, & Nord, 2007). Since empathy is stronger for in-group members, retributive motives may play a stronger role in third-party punishment when the victim is an in-group member (Lieberman & Linke, 2007). This insight has potential implications for determining the composition of juries.

Research in evolutionary game theory has examined how punishment might have evolved (Boyd, Gintis, Bowles, & Richerson, 2003; Rand, Armao IV, Nakamaru, & Ohtsuki, 2010). In most of these models, the effects of punishment operate by reducing the fitness of non-cooperators, thus making

them less plentiful in subsequent generations, rather than by reforming the behavior of non-cooperators in the current generation. These models therefore assume that one key function of punishment is to make non-cooperators worse off, which does not require their knowledge that they have been punished – akin to our Hidden punishment condition. Our finding that people are indeed willing to punish non-cooperators even when such punishment cannot serve a deterrent function thus lends psychological support to the punishment mechanism employed by evolutionary models.

Although costly punishment often has the effect of increasing cooperation (Fehr & Fischbacher, 2003; Balliet et al., 2011), whether people punish 'altruistically' in a psychological sense, with the explicit goal of promoting cooperation, remains hotly debated (Guala, 2012; McCullough, Kurzban, & Tabak, 2012; Yamagishi, Horita, & Mifune, 2012). Our results offer some resolution to this debate. We show that punishers are motivated in large part by a genuine preference to reduce the payoffs of norm violators, even in the absence of opportunities to enforce norms. Such 'hidden' punishment cannot be considered 'altruistic' because it cannot produce any social benefits. At the same time, we provide evidence that punishers have preferences for norm enforcement, in that punishers are more likely to punish, and spend more on punishment, when norms can be communicated. This could reflect an altruistic motive to deter future norm violations, or may instead reflect a retributive desire to inflict emotional harm. Regardless, the substantial contribution of retributive motives to costly punishment suggests that informal peer sanctions may not be the most efficient means of promoting

cooperation. Humans possess psychological mechanisms that can lead to destructive behavior that is sub-optimal in terms of deterring future harms. Further research is needed to understand how such motives influence the decisions of judges and juries.

**Acknowledgements**

MC is supported by a Sir Henry Wellcome Postdoctoral Fellowship from the Wellcome Trust (WT092217MA). The research was supported by the Swiss National Center of Competence in the Affective Sciences and the European Research Commission for grant number 295642, FEP, the Foundations of Economic Preferences.

**References**

Balliet, D., Mulder, L., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin, 137*, 594-615.

Batson, C. D., Kennedy, C. L., & Nord, L. A. (2007). Anger at unfairness: is it moral outrage? *European Journal of …*.

Baumgartner, T., Knoch, D., Hotz, P., Eisenegger, C., & Fehr, E. (2011). Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice. *Nature Neuroscience*, *14*(11), 1468–1474.

Bentham, J. (1970). The utilitarian theory of punishment. In J. Bentham, J. H. Burns, & H. L. A. Hart (Eds.), *An Introduction to Principles of Morals and Legislation*. London: Athlone.

Blount, S. (1995). When Social Outcomes Aren′ t Fair: The Effect of Causal Attributions on Preferences. *Organizational Behavior and Human Decision Processes*, *63*(2), 131–144.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*,*100*(6), 3531-3535.

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron*, *60*(5), 930–940.

Buckholtz, J., & al, E. (2012). The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*.

Carlsmith, K. M. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology*, *42*(4), 437–451. doi:10.1016/j.jesp.2005.06.007

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish?: Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, *83*(2), 284.

Crockett, M. J., Apergis-Schoute, A., Herrmann, B., Lieberman, M. D., Müller, U., Robbins, T. W., & Clark, L. (2013). Serotonin modulates striatal responses to fairness and retaliation in humans. *The Journal of Neuroscience*.

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(17), 6889–6892. doi:10.1073/pnas.1018033108/-/DCSupplemental/pnas.201018033SI.pdf

de Quervain, D. J. F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science; Science*.

Delgado, M. R., Li, J., Schiller, D., & Phelps, E. A. (2008). The role of the striatum in aversive learning and aversive prediction errors. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1511), 3787–3800. doi:10.1098/rstb.2008.0161

Ellsworth, P. C., & Ross, L. (1983). Public Opinion and Capital Punishment: A Close Examination of the Views of Abolitionists and Retentionists. *Crime & Delinquency*, *29*(1), 116–169. doi:10.1177/001112878302900105

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness—Intentions matter. *Games and Economic Behavior*, *62*(1), 287–303.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785–791.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87.

Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*.

Fudenberg, D., & Pathak, P. A. (2010). Unobserved punishment supports cooperation. *Journal of Public Economics*, *94*(1-2), 78–86. doi:10.1016/j.jpubeco.2009.10.007

Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment.*Philosophical Transactions of the Royal Society B: Biological Sciences*,*364*(1518), 791-806.

Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, *35*(01), 1–15. doi:10.1017/S0140525X11000069

Haber, S. N., & Knutson, B. (2009). The Reward Circuit: Linking Primate Anatomy and Human Imaging. *Neuropsychopharmacology*, *35*(1), 4–26. doi:10.1038/npp.2009.129

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., et al. (2006). Costly Punishment across Human Societies. *Science, New Series*, *312*(5781), 1767–1770.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362-1367.
Hirsch, von, A. (1986). *Doing Justice*. Westford, Mass.: Northeastern University Press.
Jensen, K. (2010). Punishment and spite, the dark side of cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1553), 2635–2650. doi:10.1111/j.1420-9101.2006.01258.x
Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, *10*(12), 1625–1633. doi:10.1038/nn2007
Kant, I. (1965). The Penal Law and the Law of Pardon. In J. Ladd (Trans.), *The Metaphysical Elements of Justice*. Indianapolis: Bobbs-Merrill.
Kurzban, R., DESCIOLI, P., & OBRIEN, E. (2007). Audience effects on moralistic punishment☆. *Evolution and Human Behavior*, *28*(2), 75–84. doi:10.1016/j.evolhumbehav.2006.06.001
Lieberman, D., & Linke, L. (2007). The effect of social category on third party punishment. *Evolutionary Psychology*, *5*(2), 289–305.
McCullough, M. E., Kurzban, R., & Tabak, B. A. (2012). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences*, 1–15. doi:10.1017/S0140525X11002160
Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231.
Piazza, J., & Bering, J. M. (2008). The effects of perceived anonymity on altruistic punishment. *Evolutionary Psychology*, *6*(3), 487–501.
Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59–63. doi:10.1016/j.tics.2005.12.004
Robbins, T. W., & Arnsten, A. F. T. (2009). The Neuropsychopharmacology of Fronto-Executive Function: Monoaminergic Modulation. *Annual Review of Neuroscience*, *32*(1), 267–287. doi:10.1146/annurev.neuro.051508.135535
Rand, D. G., & Nowak, M. A. (2011). The evolution of antisocial punishment in optional public goods games. *Nature Communications*, *2*, 434.
Rand, D. G., Armao IV, J. J., Nakamaru, M., & Ohtsuki, H. (2010). Anti-social punishment can prevent the co-evolution of punishment and cooperation.*Journal of theoretical biology*, *265*(4), 624-632.
Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, *300*(5626), 1755–1758.
Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, *303*(5661), 1157–1162.
Tunick, M. (1992). *Punishment: Theory and Practice*. Berkeley: University of California Press.
Vidmar, N., & Miller, D. T. (1979). Socialpsychological processes underlying attitudes toward legal punishment. *Law & Soc'y Rev.*, *14*, 565.
Vlaev, I. (2012). How different are real and hypothetical decisions? Overestimation, contrast and assimilation in social interaction. *Journal of*

*Economic Psychology*, *33*(5), 963–972. doi:10.1016/j.joep.2012.05.005
Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(20), 7398.
Yamagishi, T., Horita, Y., & Mifune, N. (2012). Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. Presented at the Proceedings of the …. doi:10.1073/pnas.1212126109/-/DCSupplemental
Yamagishi, T., Horita, Y., Takagishi, H., Shinada, M., Tanida, S., & Cook, K. S. (2009). The private rejection of unfair offers and emotional commitment. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(28), 11520–11523.

**Table and Figure Captions**

**Table 1.** Different punishment motives predict different patterns of punishment across experimental conditions.

| Punishment motive | Prediction |
|---|---|
| Deterrence | Open Unfair > Fair<br>Hidden Unfair = Fair<br>Computer Unfair = Fair |
| Retribution | Open Unfair > Fair<br>Hidden Unfair > Fair<br>Computer Unfair = Fair |
| Payoff-based (e.g., spite, inequality aversion) | Open Unfair > Fair<br>Hidden Unfair > Fair<br>Computer Unfair > Fair |

**Figure 1. Experimental design.** Each trial consisted of three stages. In the *trust* stage, the punisher (P) and bystander (B) entrust their endowments to the trustee (T). In the *back-transfer* stage, P and B receive back-transfers from T. In the *punishment* stage, P decides whether to punish T. We varied the back-transfer mechanism across three experimental conditions. (A) In second-party punishment trials, T decides how much to send back to P, while the computer decides how much T sends back to B. Thus, P's punishment decision concerns T's intentional back-transfer towards P. (B) In third-party punishment trials, T decides how much to send back to B, while the computer decides how much T sends back to P. Thus, P's punishment decision concerns T's intentional back-transfer towards B. (C) In computer control

trials, the computer decides how much T sends back to both P and B. Thus, P's punishment decision concerns only the payoff differences between players.

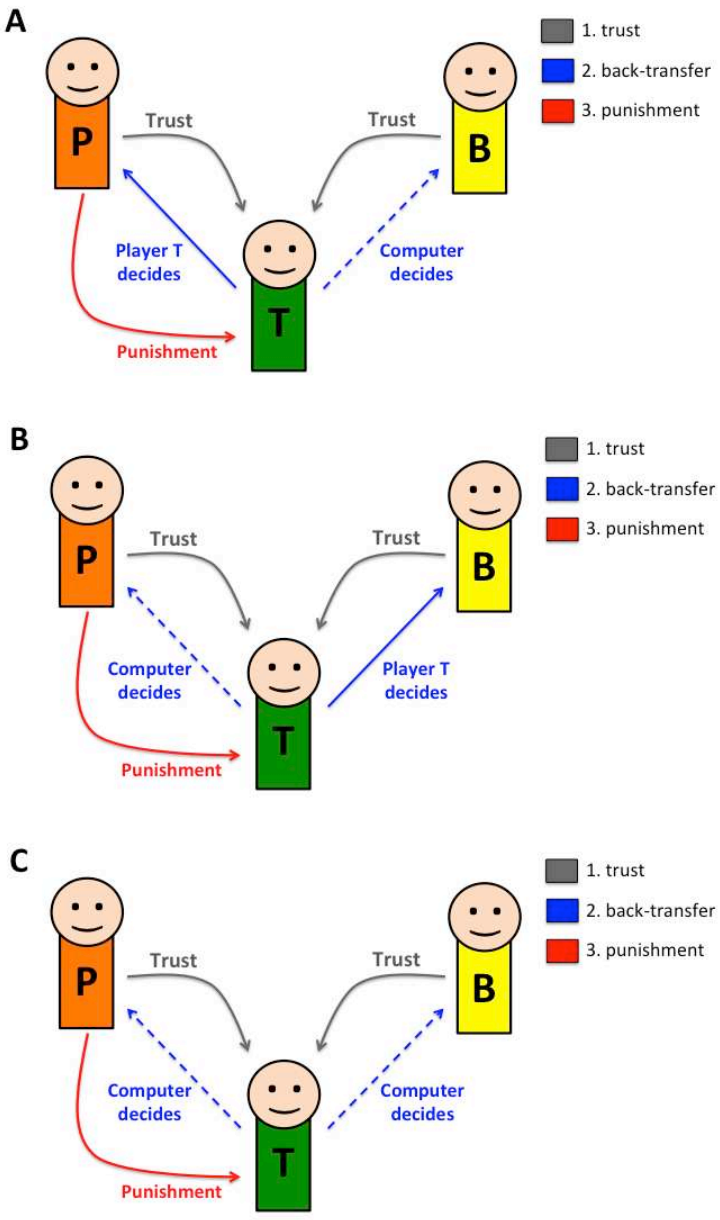

**Figure 2. Retribution and deterrence in second- and third-party punishment.** Punishment likelihoods (A) and mean amount spent (B) for second-party punishment (2PP; red) and third-party punishment (3PP; blue), in the open (lined) and hidden (solid) conditions. Error bars depict SEM.

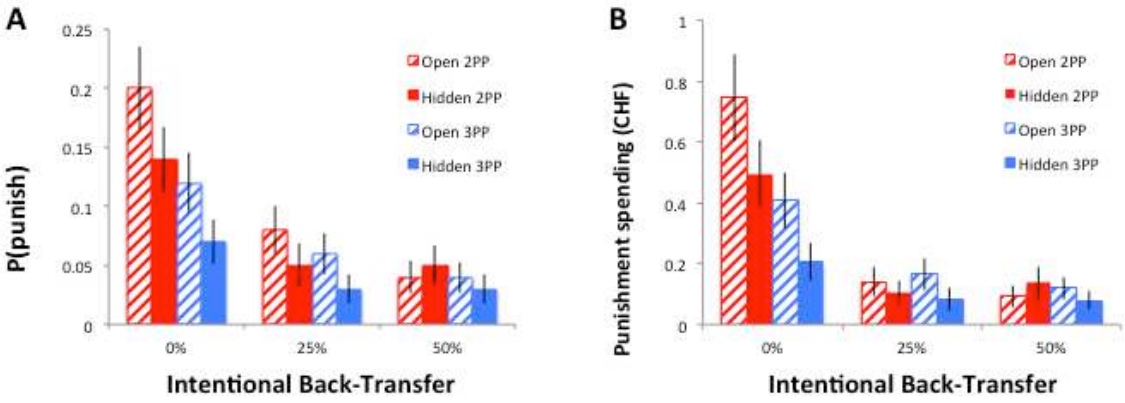

**Figure 3. Retribution and payoff-based motives in second- and third-party punishment.** Punishment likelihoods (A) and mean amount spent (B) for hidden punishment levels when back-transfers resulted from intentional decisions by trustees (solid) versus when back-transfers resulted from the computer's decision (lined), in the second-party punishment (2PP; red) and third-party punishment (3PP; blue) conditions. Error bars depict SEM.