When and why do third parties punish outside of the lab? A cross-cultural recall study

Eric J. Pedersen[1*], William H. B. McAuliffe[2], Yashna Shah[2,3], Hiroki Tanaka[4], Yohsuke Ohtsubo[5], & Michael E. McCullough[2, 6*]

1. Department of Psychology and Neuroscience, University of Colorado Boulder, Boulder, CO, United States.

2. Department of Psychology, University of Miami, Coral Gables, FL, United States

3. Independent Researcher

4. Brain Research Institute, Tamagawa Univeristy, Tokyo, Japan

5. Department of Psychology, Kobe University, Kobe, Japan

6. Department of Psychology, University of California San Diego, La Jolla, CA, United States

*Corresponding author

Word count: 4,996

Author Note

We thank Ayumi Masuchi (Hokkai Gakuen University), Sayaka Suga (Keio University), Nobuko Asai (Kyoto Bunkyo University), Daisuke Nakanshi (Hiroshima Shudo University), Tomoko Oe (Teikyo University), and Keiko Ishii (Nagoya University) for their generous assistance with data collection. This research was supported by grants from the Air Force Office of Scientific Research (Award #FA9550-12-1-0179) to M.E.M, the John Templeton Foundation (29165) to M.E.M, and by the Expanding the Science and Practice of Gratitude Project run by UC Berkeley's Greater Good Science Center in partnership with UC Davis with funding from the John Templeton Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Correspondence concerning this article should be addressed to Eric J. Pedersen (eric.j.pedersen@colorado.edu).

Author Contributions

E.J.P., M.E.M., W.H.B.M., Y.S., and Y.O. conceived of the study. E.J.P., W.H.B.M., Y.S., H.T., and Y.O. managed data collection. E.J.P. and W.H.B.M analyzed the data and interpreted results. E.J.P. drafted the manuscript. W.H.B.M., M.E.M., Y.S., H.T., and Y.O. revised the manuscript.

Abstract

Punishment can reform uncooperative behavior, and hence could have contributed to humans' ability to live in large-scale societies. Punishment by unaffected third parties has received extensive scientific scrutiny because third parties punish transgressors in laboratory experiments on behalf of strangers that they will never interact with again. Often overlooked in this research are interactions involving people who are *not* strangers, which constitute many interactions beyond the laboratory. Across three samples in two countries (US and Japan; $N = 1{,}294$), we found that third parties' anger at transgressors, and their intervention and punishment on behalf of victims, varied in real-life conflicts as a function of how much third parties valued the welfare of the disputants. Punishment was rare (1-2%) when third parties did not value the welfare of the victim, suggesting that previous economic game results have overestimated third parties' willingness to punish transgressors on behalf of strangers.

Keywords: third-party punishment; anger; cooperation; bystander intervention

Introduction

In laboratory experimental games, the majority of third parties are willing to pay costs to punish a transgressor who has harmed a stranger, a finding that has been replicated across several cultures (Bernhard, Fischbacher, & Fehr, 2006; Henrich et al., 2005, 2006; cf. Marlowe, 2009). For instance, in an anonymous, one-shot experimental economic game called the third-party punishment game, Fehr and Fischbacher (2004) found that approximately two-thirds of third parties punished transgressions who unfairly split a sum of money. Henrich and colleagues (2006) found similar results in 15 diverse societies. Despite considerable between-society variation, subjects in all 15 societies evinced a penchant for punishment in the third-party punishment game. However, the ethnographic literature and some field studies indicate that everyday punishment might be substantially rarer than the results of these experiments imply (Balafoutas & Nikiforakis, 2012; Balafoutas, Nikiforakis, & Rockenbach, 2014, 2016; Guala, 2012). For example, Balafoutas and colleagues reported punishment rates of littering from 4% to 17% in their field studies. Furthermore, most studies have focused on simply demonstrating that third-party punishment occurs on behalf of strangers. Thus, we do not know whether (and, if so, how) people regulate their decisions about third-party punishment based on their relationships with the harmdoer and the victim. Our goals here were (a) to test whether the types of interactions modeled in the third-party punishment game also result in frequent punishment in everyday life and (b) to shed light on the decision-making systems responsible for third-party punishment.

Punishment is inherently costly to the punisher because it requires time and energy, and because it can provoke retaliation. As such, an evolutionary perspective

implies that third-party punishment is selectively employed in situations in which, on average, the cost of punishment is outweighed by the benefits to the punisher. Some researchers have proposed that a primary function of third-party punishment is to *deter aggressors from harming individuals with whom the punisher shares a fitness interest* (Pedersen, McAuliffe, & McCullough, 2018). On this view, the benefits of third-party punishment can offset its costs by deterring future harms toward victims in whom the punisher has a welfare stake (e.g., kin, mates, friends, coalition members; see also Hofmann, Brandt, Wisneski, Rockenbach, & Skitka, 2018; Lieberman & Linke, 2007). If a third party's own welfare is sufficiently interdependent with a victim's welfare, a harm imposed on the victim also *indirectly* harms the third party. We propose that third-parties' perceptions of indirect harms trigger anger toward transgressors and possibly lead to punishment. Importantly, if a third party's own welfare is sufficiently interdependent with that of the transgressor, and the transgressor benefits from the harm he or she imposes on the victim, then the third party indirectly *benefits* from the transgression. Thus, for anger to result, and possibly lead to punishment, a third party needs both a sufficiently high estimate of welfare interdependence with the victim and a sufficiently low estimate of welfare interdependence toward the transgressor, relative to the costs and benefits incurred by each.

Several lines of evidence support these hypotheses, including ethnographic evidence of punishment on behalf of genetic relatives (Boehm, 1987; Chagnon & Bugos, 1979; Ericksen & Horton, 1992); a study of violent criminals showing virtually no punishment on behalf of strangers but substantial amounts on behalf of friends, family members, and fellow gang members (Phillips & Cooney, 2005); and social psychology

experiments indicating that (a) people experience moral outrage only when self-relevant concerns are present (Batson, 2015) and (b) third parties punish on behalf of their friends but not on behalf of strangers (Pedersen, Kurzban, & McCullough, 2013; Pedersen et al., 2018). To regulate behavior in accordance with welfare interdependence considerations, humans may possess psychological systems that estimate their interdependence with others from a variety of fitness-relevant inputs. A proposed output of such a system is called a *welfare trade-off ratio* (WTR), which is hypothesized to be an internal regulatory variable that weights the welfare of another individual relative to the self and guides behavior accordingly through its effects on motivational systems (Tooby, Cosmides, Sell, Lieberman, & Sznycer, 2008). The WTR a third party holds for a victim is the ratio of the third party's valuation of the victim's welfare relative to the third party's *own* welfare, which is one (see also Balliet, Tybur, & Van Lange, 2017; Brown & Brown, 2006; Rachlin, 2015; Roberts, 2005). For example, when the third party's WTR for the victim is 0, the third party has no regard for the victim's welfare (i.e., he or she would not incur any cost to benefit the victim); when the third party's WTR for the victim is 1, the third party regards the victim's welfare as equivalent to the third party's (i.e., he or she would incur any cost outweighed by the benefit to the victim); when the third party's WTR for this victim is greater than 1, the third party regards the victim's welfare greater than his or her own. WTRs are likely updated in response to new information, such as increased confidence that the other person is either a cooperative or exploitative partner.

Our proposal is that anger and third-party punishment are triggered when a third party perceives that he or she has incurred a net cost because of a harm to the victim. A

perceived net cost to the third party would occur when the cost to the victim, discounted by the third party's WTR for the victim, exceeds the benefit to the transgressor, discounted by the third party's WTR for the transgressor. The decision-making system(s) that weigh this information must also integrate information about the costs and benefits associated with punishment, such as the likelihood that the transgressor retaliates against the punisher and the likelihood that the punishment successfully deters future costs to the punisher via transgressions against the same victim. As others have suggested, third-party punishers can also possibly benefit from enhanced reputations (Barclay, 2006; Jordan, Hoffman, Bloom, & Rand, 2016; Kurzban, DeScioli, & O'Brien, 2007) and the deterrence of future direct harms to the punisher (Krasnow, Delton, Cosmides, & Tooby, 2016). We think all of these data might be integrated into a decision-making system that leads third parties to punish when, on average, the circumstances of the situation are (or ancestrally were) correlated with fitness benefits to the punisher. However, holding everything else equal, the third party's WTR for the victim should predict more anger toward the transgressor and intervention on behalf of the victim, and the third party's WTR for the transgressor should predict less anger toward the transgressor and intervention on behalf of the victim.

Here, we sought to address two limitations to previous studies by using a recall method about people's actual experiences. First, most studies have been conducted via experiments in which interactions are anonymous and one-shot, which might lack generalizability because, outside of circumscribed situations that were not present ancestrally, people do not interact with others anonymously. Thus, to punish someone typically means to encounter them, and the interaction can be witnessed by others.

The second limitation of existing experiments is that they limit how third parties can respond: Often, they are restricted between either punishing the transgressor or doing nothing, which they are explicitly prompted to choose between. In real life, many actions can be taken in response to a transgression: third parties can *help* the transgressor, they can mediate, and they can recruit help (e.g., call the police, gather bystanders). Thus, laboratory-based estimates of third parties' willingness to punish might be inflated because punishing is (a) often the only mechanism for action available and (b) explicitly prompted by the design of the experiment. Experiments suggest that adding options such as compensating a victim or rewarding a transgressor causes punishment to decline as compared to standard designs (Chavez & Bicchieri, 2013; Ohtsubo, Sasaki, Nakanishi, & Igawa, 2018; Pedersen et al., 2013). Indeed, ethnographic accounts of small-scale societies suggest that people rarely unilaterally punish in the real world unless they or their kin suffer a serious harm (Guala, 2012). Hence, what appears to be a penchant for third-party punishment in some experiments may result from subjects' lack of access to preferable alternatives. To study how possessing multiple alternatives for responding to a conflict affects rates of punishment, here we examined third-party punishment alongside other interventions on behalf of the victim (i.e., getting involved in the conflict to stop the harm to the victim without imposing obvious costs on the transgressor). Because intervention should generally be less costly to the third party than punishment, we predicted that intervention would be more common than punishment. To sum up, we predicted that a third party's WTR for the victim would positively predict intervention, punishment, and anger toward the

transgressor, and that the third party's WTR for the transgressor would negatively predict intervention, punishment, and anger toward the transgressor.

We also attempted to complement existing studies by using a design that increases generalizability. Field studies on third-party punishment have observed how people react to specific transgressions (e.g., littering) in specific locations (e.g., the subway) with specific populations (e.g., adults in Cologne). We cannot infer the overall rate of third-party punishment, aggregated across situations that vary in factors that promote or inhibit action, from how much punishment occurs in any one situation. Hofmann et al. (2018) increased generalizability by having participants report their desires to punish in an experience sample but they did not measure behavior. Here we sought also to increase generalizability across populations by recruiting an undergraduate samples in the United States and Japan, as well as an adult sample in the United States.

Method

Subjects

US Students: 619 University of Miami undergraduates participated for partial course credit. US MTurkers: 649 subjects from Amazon Mechanical Turk (restricted to US users) participated for $0.25. Japanese students: 416 undergraduates from seven universities in various regions of Japan participated for partial course credit. Sample sizes were determined by maximal recruitment efforts for our time and budget constraints, and no analyses were conducted before data collection was complete.

Procedure

The student samples completed the study via pencil and paper; MTurkers participated online (procedure and measures were identical except where noted). After providing consent, subjects were given a questionnaire in which they were asked to "think of the last situation you can recall in which you witnessed someone attack, insult, or otherwise mistreat another person." Subjects described the conflict in one sentence and then asked whether they "help[ed] either person in any way" and, if so, to describe what they did. These free responses were coded independently by two raters for whether the subject's response was "punishment," "intervention," or "nothing" (US samples initial agreement: 94%, $\kappa = .89$; Japanese sample initial agreement: 95%, $\kappa = -.02$; The kappa for the Japanese sample was poor despite the high agreement due to initial disagreements on how the small number of intervention and punishment responses [20 total] should be coded; the vast majority [396] of responses in the Japanese sample were coded as "nothing."). Differences in coding between raters were resolved with discussion prior to any analyses and erred on the side of coding for action rather than inaction. We defined punishment as any imposition of costs on the transgressor during the conflict, which could be either physical (e.g., hitting, tackling, aggressive pushing) or verbal (e.g., yelling, insulting). We defined intervention as any action that involved the subject during the conflict that was not punishment. For example, physically separating the disputants, verbally sticking up for the victim, or calling the police were coded as intervention. For our purposes here (i.e., to capture interactions similar to those modeled by the third-party punishment game), we are concerned with third-party intervention on behalf of the victim *during* the conflict. Thus,

all other responses, including helping the transgressor, were coded as “nothing.” Consequently, we do not infer that people who did nothing under our coding system were necessarily apathetic to the plight of the victim. For example, we coded consoling the victim after the conflict took place as “nothing,” despite the fact that consolation typically reflects caring. Moreover, some participants may have wanted to intervene, but did not do so out of fear of retaliation or other countervailing considerations.

Exclusion criteria. We excluded from analyses cases in which the subject reported a conflict they either did not directly witness or that contained strong physical or societal barriers to intervention (such as being in a separate car or building, witnessing a boss aggress against a coworker, witnessing a parent scold a child; see SOM text and Tables S1, S9, S10 for all exclusion criteria, more details, and exploratory analyses) so as to not bias our analyses against finding intervention and punishment. These exclusions were made prior to all data analyses and led to sample sizes of: 463 US students (75% of total N); 448 US MTurkers (69%); 383 Japanese students (92%).

Measures

Welfare trade-off ratios (WTRs). WTRs were calculated both for the transgressor and the victim. Subjects were asked to imagine that the experimenters could either pay a sum of money to themselves or the focal individual from the conflict. Then, subjects made a series of ten binary decisions that required them to indicate whether they would choose to take an amount of money for themselves (ranging from $0 to $85) or direct a fixed amount of money ($75) to the focal individual from the conflict (Rachlin, 2015). We calculated WTR values by finding the indifference (or switch) point on the scale, defined as the average between the last amount the participant selected for him/herself and the

first amount forgone to give $75 to the other person, divided by $75. For instance, if a subject chose $55 for themselves (instead of $75 for the other person) and in the next decision chose $75 for the other person (instead of $45 for themselves), the indifference point was the midpoint between $55 and $45 ($50). WTRs were calculated by taking the ratio of the indifference point relative to the fixed amount (in this example, WTR = 50/75, or 0.67). The scale used here gives a possible WTR range of 0.00 to 1.13. We did not calculate WTR scores for subjects who provided more than 1 switch-point over the course of the 10 decisions, or for those who first selected $75 for the other person (versus $85 for themselves) and then subsequently switched to selecting money for themselves. Scores for these cases were treated as missing values and thus subjects were deleted listwise when analyses contained variables on which the subjects were missing (missing WTR values: US students [transgressor: 6; victim: 18], US MTurkers [transgressor: 11; victim: 19] Japan [transgressor: 0; victim: 1]).

Relationship with the disputants. Subjects reported how each party was related to them (e.g., family member, friend, acquaintance, or stranger). See SOM Figures S1 and S2 for WTR distributions by relationship category, which illustrate that WTRs track welfare interdependence and provide more fine-grained information than does relationship category by itself (see Figures S3 and S4 for third party response by relationship category).

Emotional reactions. Subjects rated how angry and empathic they felt toward both the transgressor and the attacked person on six-point rating scales with response options ranging from 0 (not at all) to 5 (extremely). Subjects also rated the extent to

which they liked the transgressor and the attacked person at the time of the attack on a scale from -3 (*disliked very much*) to 3 (*liked very much*).

Data Availability. All data are available at: https://osf.io/zmrxf/ and all syntax used for the analyses are available at: https://osf.io/wgbvd/.

Results

All analyses were conducted in R version 3.4.4. We treated subjects' responses to the conflict (nothing, intervene, punish) as an ordinal outcome and analyzed it using ordinal logistic regression with the VGAM package. We ran an initial model with WTR for the victim (WTR_{victim}), WTR for the transgressor ($WTR_{transgressor}$), and dummy codes for sample (US Students and US MTurkers; JPN students were the reference group) as predictors, along with all of their interactions (Table S2). WTR_{victim} and the sample dummies were significant predictors in this model, whereas $WTR_{transgressor}$ was not. None of the slopes significantly varied as a function of sample, nor was there an interaction between WTR_{victim} and $WTR_{transgressor}$. To verify that the lack of an effect of transgressor WTR was not due to the presence of non-significant interactions, we dropped the interaction terms and re-ran the model with WTR_{victim}, $WTR_{transgressor}$, and sample as predictors. $WTR_{transgressor}$ was not significant in this reduced model ($b = -.26$, $p = .211$), but the WTR_{victim} and sample terms were (ps <.001; see Table S3). For simplicity in interpreting the effects of WTR_{victim} and sample, we ran a final model including only those terms (pseudo R^2 relative to intercept-only model = .158). Results of this model are displayed in Table 1 and the model-predicted probabilities of third-parties' response to the conflict are displayed in Figure 1.

Table 1. Ordinal logistic regression model results predicting response to conflict as a function of WTR_{victim} and sample.

Parameter	b	95% CI	OR	p
Intercept 1	-2.11	[-2.43, -1.80]	0.12	< .001
Intercept 2	-5.18	[-5.66, -4.70]	0.01	< .001
WTR_{victim}	1.17	[0.87, 1.46]	3.21	< .001
US MTurkers	0.89	[0.54, 1.23]	2.42	< .001
US Students	1.15	[0.82, 1.48]	3.17	< .001

Note. Intercept 1 refers to the log odds of responding with intervention or punishment, relative to responding with doing nothing. Intercept 2 refers to the log odds of responding specifically with punishment, relative to responding with intervention or doing nothing. WTR_{victim} (welfare trade-off ratio toward the victim) is a continuous predictor ranging from 0-1.13; sample variables are dummy codes. Reference group = JPN Students. Pseudo R^2 relative to intercept-only model = .158.

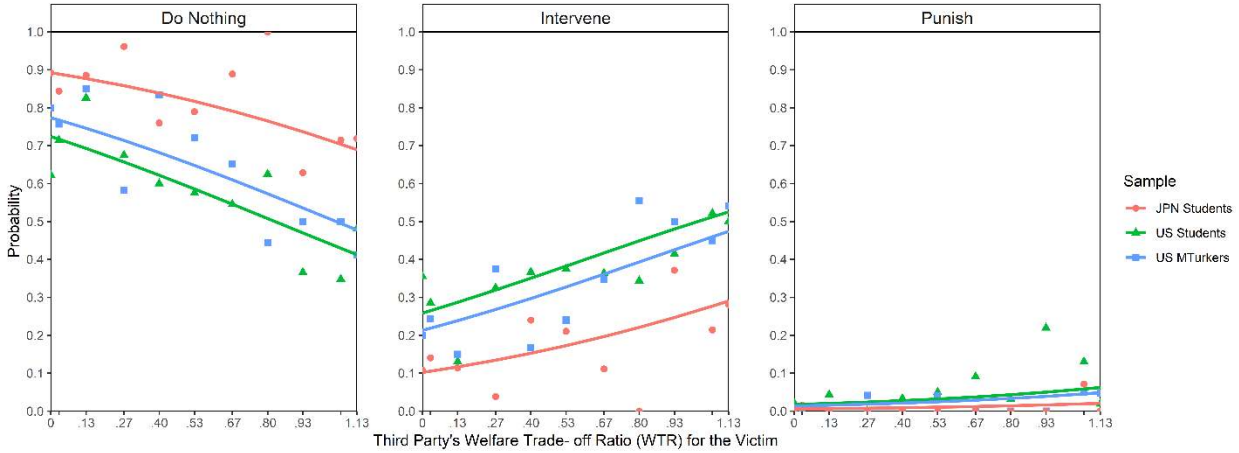

Figure 1. Third party's response to conflict. Solid lines represent model-predicted probabilities; points represent raw proportions.

The association of WTR_{victim} with subjects' likelihood of intervening was significant (b = 1.17, 95% CI: [0.87, 1.46], p < .001). Exponentiating this coefficient results in an odds ratio of 3.21, indicating that a one unit increase on the WTR scale (which ranged from 0 – 1.13) leads to a 221% increase in the likelihood of taking an action on behalf of the victim (intervening or punishing). A likelihood ratio test indicated that dropping the proportional odds assumption of ordinal logistic regression did not significantly improve model fit, X^2 (3) = 5.35, p = .148. Thus, the 221% increase as a

function of WTR_{victim} is most parsimoniously applied to both the odds of either intervening or punishing, relative to doing nothing, and to the odds of punishing, relative to doing nothing or intervening.

Despite this large effect of WTR_{victim}, punishment was infrequent. Indeed, the model-predicted probabilities of punishing the transgressor when WTR_{victim} was 0 ranged from only .01-.02. At the maximum value for WTR_{victim} (1.13), model-predicted probabilities of punishment ranged from (.02-.06). In contrast, intervention was more common: model-predicted probabilities of intervention when WTR_{victim} was 0: .10-.26; when WTR_{victim} was 1.13: .29-.52).

As indicated above, the effect of WTR_{victim} did not vary by sample. However, both US Students (OR = 3.17, $p < .001$) and US MTurkers (OR = 2.42, $p < .001$) were significantly more likely to intervene or punish than were Japanese students. US MTurkers were marginally less likely to intervene or punish than were US Students (OR = .77, $p = .074$).

We ran two additional ancillary analyses. First, to examine whether the type of the conflict affected the likelihood of intervention and punishment, we added a dummy-coded predictor for whether the conflict was physical or verbal, as well as the interaction of this with sample (Table S4). There was a significant effect (b = .60, OR = 1.81, $p = .036$) indicating that third parties were 80% more likely to intervene or punish when the conflict was physical than when it was verbal (there were no significant interactions with sample, $ps > .30$).

Second, researchers have posited that punishment occurs in response to the violation of a social norm for the purposes of upholding group cooperative norms,

regardless of who the target of the transgression is (for review, see Krasnow, Cosmides, Pedersen, & Tooby, 2012). In our US MTurker and Japanese student samples, we asked participants whether the transgressor's "actions [broke] a social norm or rule for how people should treat each other" (US MTurkers: 70% of cases were social norm violations; JPN Students: 49%). We added a dummy-coded predictor for social norm violation, as well as the interaction with sample, to the model (separately from the type of conflict model reported above; see Table S5 for full results). Neither the dummy code nor the interaction were significant (ps > .11). However, removing the nonsignificant interaction from the model revealed a significant effect for social norm violation (b = .40, OR = 1.49, p = .037), indicating that third parties were about 49% more likely to intervene when they perceived that a social norm had been violated. Importantly, when controlling for social norm violation in the model, the effect of WTR_{victim} remained unchanged (b = 1.24, OR = 3.44, p < .001). These results are consistent with the idea that social norm violations can provoke intervention independent of welfare interdependence considerations.

We analyzed self-reported anger toward the transgressor (scale: 0-5) with linear regression. We ran an initial model with WTR_{victim}, $WTR_{transgressor}$, and sample (dummy-coded) as predictors, along with all their interactions (Table S6). No terms that included the interaction between WTR_{victim} and $WTR_{transgressor}$ were significant. Thus, we dropped all these terms and re-ran the model with WTR_{victim}, $WTR_{transgressor}$, sample, and their interactions as predictors. Results of this model are displayed in Table 2.

Table 2. Linear regression model results predicting anger toward transgressor as a function of WTR_{victim}, $WTR_{transgressor}$, and sample

	Reference Group = JPN Students			Reference Group = US Students		
Parameter	b	95% CI	p	b	95% CI	p
Intercept	2.48	[2.27, 2.70]	<.001	1.92	[1.66, 2.18]	<.001
WTR_{victim}	1.42	[1.08, 1.76]	<.001	1.78	[1.44, 2.12]	<.001
US MTurkers	0.10	[-0.24, 0.44]	0.560	0.66	[0.29, 1.04]	<.001
US Students	-0.56	[-0.90, -0.22]	0.001	-	-	-
JPN Students	-	-	-	0.56	[0.22, 0.90]	0.001
$WTR_{transgressor}$	-2.00	[-2.54, -1.47]	<.001	-0.77	[-1.24, -0.29]	0.002
WTR_{victim} *US MTurkers	-0.24	[-0.73, 0.25]	0.335	-0.60	[-1.09, -0.12]	0.015
Transgressor WTR*US MTurkers	0.45	[-0.25, 1.15]	0.209	-0.78	[-1.44, -0.13]	0.019
WTR_{victim} *US Students	0.36	[-0.12, 0.84]	0.139	-	-	-
$WTR_{transgressor}$ *US Students	1.23	[0.52, 1.95]	<.001	-	-	-
WTR_{victim} *JPN Students	-	-	-	-0.36	[-0.84, 0.12]	0.139
$WTR_{transgressor}$ *JPN Students	-	-	-	-1.23	[-1.95, -0.52]	<.001

Note. WTR = welfare trade-off ratio; WTR_{victim} and $WTR_{transgressor}$ are continuous predictors ranging from 0-1.13; sample variables are dummy codes. The two models are identical but recoded with different reference groups.

WTR_{victim} significantly predicted greater anger toward transgressors within each of the three samples (bs = 1.17 to 1.78, ps < .001). Consistent with our prediction, though unlike the intervention and punishment results, $WTR_{transgressor}$ significantly predicted less anger toward transgressors within each of the three samples (bs -.77 to -2.00, ps <.003). There were some differences amongst data sets in the magnitude of these effects, but not in kind (i.e., they did not change in statistical significance or direction; see Table 2). We ran similar ancillary analyses with conflict type and social norm violation predicting anger as we did for intervention and punishment (see Tables S7 and S8 for full results). Unlike for intervention and punishment, conflict type (physical vs. verbal) had no effect on anger toward transgressors (b = -.06, p = .720). However, as they did for intervention and punishment, social norm violations predicted anger toward transgressors (b = 1.06, p < .001) without substantively changing the effects of WTR_{victim} and $WTR_{transgressor}$.

Discussion

Here we proposed that a major function of third-party punishment is to deter aggressors from harming individuals with whom the punisher shares a fitness interest, and that the psychological mechanisms that regulate punishment take into account the punisher's perceived welfare interdependence with the disputants in a conflict (Pedersen et al., 2018). To test these hypotheses, we asked U.S. students, U.S. Mechanical Turk workers, and Japanese students to recall how they responded the last time they observed a conflict. The recall study method ensures a wide sampling of situations, and thus high generalizability to real-life conflicts. We found that third parties' WTRs for the victim in a conflict indeed predicted anger, intervention, and punishment on behalf of the victim. We also found that third parties' WTRs for the transgressor were *negatively* associated with anger toward the transgressor, but not with intervention or punishment as we had predicted. Besides the possibility that WTR for the transgressor truly does not predict intervention and punishment, one possibility for the lack of these associations is that third parties who intervene or punish may temporarily hold a negative WTR for the transgressor—that is, they are willing to incur costs to inflict costs. Because our WTR scale only went down to zero, any negative WTRs would have manifested as zeros and thus the variability of the scale could have been restricted (see Figure S2, which suggests this may have been the case), which would limit our power to detect an effect.

These findings were generally consistent across our three samples and never differed in kind, only magnitude. For intervention and punishment, the effect of third parties' WTR for the victim was constant across all samples, though Japanese students intervened and punished less often than either US sample. For anger, there were minor

differences among the samples in the magnitude of the effects of third parties' WTRs for the victim and the transgressor, but they remained in the same, predicted directions in all samples. Thus, we have initial evidence that our findings are at least somewhat generalizable beyond a US student population, both to a more general US population and to Japanese students.

The low model-predicted probabilities of punishment (≤.02) we found when WTR for the victim was 0 suggest that the frequency of third-party punishment has likely been overstated in the literature that has focused on results from laboratory-based experimental economics games (for similarly low rates of punishment in naturalistic settings, see Balafoutas et al., 2014, 2016). Thus, in addition to providing support for our hypotheses that third-party anger, intervention, and punishment vary as a function of the prospective punisher's WTRs toward disputants in a conflict, the present study adds to a growing body of evidence suggesting that direct third-party punishment on behalf of strangers is not a common feature of human cooperation (Guala, 2012; Krasnow et al., 2012, 2016; Kriss, Weber, & Xiao, 2016; Pedersen et al., 2013, 2018; Phillips & Cooney, 2005).

We recognize that some might view our design choices here as restrictive because we limited our scope to conflicts where there was a direct harm to a victim and only considered intervention and punishment that occurred in the moment. These were intentional choices to mimic the types of interactions that are created in the third-party punishment game (Fehr & Fischbacher, 2004), which typically shows that a majority of people anonymously engage in immediate, uncoordinated, costly punishment on behalf of victims. These findings have been generalized to draw conclusions about humans'

willingness to directly punish transgressions and what this implies for the evolution of cooperation in humans (Fehr & Fischbacher, 2003; Henrich et al., 2010, 2006; for review, see Pedersen et al., 2018). Our results suggest that people are much less likely to engage in this type of punishment than a direct generalization of previous laboratory experiments would imply, though perhaps future studies will show higher rates of after-the-fact punishment with low-WTR parties than we found here. Thus, it is important to note that our data cannot speak directly to a broader range of social norm violations, some of which could be likely to evoke punishment. Additionally, we did not focus on indirect types of retaliation, such as gossip, or other mechanisms that are likely important to maintaining cooperation and social norms, such as partner choice.

Additionally, the higher rate of intervention than punishment we observed here comports well with evidence suggesting that people prefer alternatives to punishment (e.g. helping the victim) when they are available (Balafoutas et al., 2014, 2016; Chavez & Bicchieri, 2013). It also suggests that shifting focus beyond punishment could be a fruitful approach to more fully understanding how third parties respond to conflicts in the real world. We do note that the amount of reported intervention could have been inflated due to our asking subjects to report whether they had “helped” either person involved in the conflict, though this was asked after subjects had already chosen a particular conflict to recall and thus probably did not bias the choice of event in the first place. It is also possible that our prompt elicited different recollections between the US and Japanese samples, which could explain the difference intervention and punishment rates between the countries.

This study had some limitations. First, memory limitations may have prevented people from accurately recalling the details of past events. For example, subjects' WTRs for the victims and transgressors were retrospective; consequently, they might have been disproportionately reflective of their *current* WTRs for the victims and transgressors. Indeed, it is possible that choosing to intervene or punish increased subjects' commitment toward victims, and thus could have increased their WTRs. Though we cannot rule this possibility out given the nature of our data, we do note that recalled WTRs varied expectedly as a function of the relationship between subjects and the victims (see SOM), which suggests that reported WTRs did at least moderately correspond to the existing relationships.

Second, subjects' reports might have been distorted by socially desirable responding. The low levels of punishment speak against this concern, but it might have played a role in intervention responses. The possibility of socially desirable responding in combination with our exclusion of cases *a priori* from situations in which the costs of intervening were very steep (e.g., conflicts involving guns, multiple transgressors), leads us to believe that the current study did not underestimate intervention and punishment frequency. Finally, we did not code for consolation—attempting to make the victim feel better after the conflict had ended—and instead treated it as the same as doing nothing because it had no material effect on the conflict as it was occurring. Although consolation is certainly a much less costly helping behavior, it nevertheless may help the victim and is an important area for future research (De Waal, 2008).

To conclude, the present investigation moved beyond the question, "do people punish on behalf of strangers," to ask, "when and why do people intervene on behalf of

others?" Our method sampled intervention and punishment decisions across a wide range of situations and multiple populations, complementing studies that have examined punishment (and the desire to punish) in specific real-life situations (Balafoutas et al., 2014, 2016; Hofmann et al., 2018). Our results converged with results from these other studies, suggesting that intervention is much more common than punishment in everyday life. Perceived welfare interdependence with the victim emerged as the strongest predictor of intervention and punishment, signaling its promise as an explanation of involvement of others' affairs.

References

Balafoutas, L., & Nikiforakis, N. (2012). Norm enforcement in the city: a natural field experiment. *European Economic Review*, *56*(8), 1773–1785.

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

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2016). Altruistic punishment does not increase with the severity of norm violations in the field. *Nature Communications*, *7*.

Balliet, D., Tybur, J. M., & Van Lange, P. A. M. (2017). Functional interdependence theory: An evolutionary account of social situations. *Personality and Social Psychology Review*, *21*(4), 361–388.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*, 325–344.

Batson, C. D. (2015). *What's wrong with morality?: a social-psychological perspective*. Oxford University Press.

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature*, *442*, 912–915.

Boehm, C. (1987). *Blood revenge: The enactment and management of conflict in Montenegro and other tribal societies* (2nd ed.). Philadelphia: University of Pennsylvania Press.

Brown, S. L., & Brown, R. M. (2006). Selective investment theory: Recasting the functional significance of close relationships. *Psychological Inquiry*, *17*(1), 1–29.

Chagnon, N., & Bugos, P. (1979). Kin selection and conflict: An analysis of a

Yanomamö ax fight. (N. Chagnon & W. Irons, Eds.), *Evolutionary Biology and Human Social Behavior: An Anthropological Perspective*. North Scituate: Duxbury.

Chavez, A. K., & Bicchieri, C. (2013). Third-party sanctioning and compensation behavior: Findings from the ultimatum game. *Journal of Economic Psychology*, *39*, 268–277.

De Waal, F. B. M. (2008). Putting the altruism back into altruism: the evolution of empathy. *Annu. Rev. Psychol.*, *59*, 279–300.

Ericksen, K. P., & Horton, H. (1992). “Blood feuds”: Cross-cultural variations in kin group vengeance. *Behavior Science Research*, *26*, 57–85.

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*, 63–87.

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

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & al., et. (2005). “Economic man” in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, *28*, 795–855.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., … Ziker, J. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science*, *327*, 1480–1484.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., … Ziker, J. (2006). Costly punishment across human societies. *Science*, *312*, 1767–1770.

Hofmann, W., Brandt, M. J., Wisneski, D. C., Rockenbach, B., & Skitka, L. J. (2018). Moral punishment in everyday life. *Personality and Social Psychology Bulletin*. https://doi.org/10.1177/0146167218775075

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

Krasnow, M. M., Cosmides, L., Pedersen, E. J., & Tooby, J. (2012). What Are Punishment and Reputation for? *PLoS One*, *7*(9), e45662.

Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit. *Psychological Science*.

Kriss, P. H., Weber, R. A., & Xiao, E. (2016). Turning a blind eye, but not the other cheek: on the robustness of costly punishment. *Journal of Economic Behavior & Organization*.

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*, 75–84.

Lieberman, D., & Linke, L. (2007). The effect of social category on third party punishment. *Evolutionary Psychology*, *5*, 289–305.

Marlowe, F. W. (2009). Hadza cooperation. *Human Nature*, *20*(4), 417–430.

Ohtsubo, Y., Sasaki, S., Nakanishi, D., & Igawa, J. (2018). Within-individual associations among third-party intervention strategies: Third-party helpers, but not punishers, reward generosity. *Evolutionary Behavioral Sciences*, *12*(2), 113–125.

Pedersen, E. J., Kurzban, R., & McCullough, M. E. (2013). Do humans really punish altruistically? A closer look. *Proceedings of the Royal Society B: Biological*

Sciences, *280*(1758).

Pedersen, E. J., McAuliffe, W. H. B., & McCullough, M. E. (2018). The unresponsive avenger: More evidence that disinterested third parties do not punish altruistically. *Journal of Experimental Psychology: General*, *147*(4), 514.

Phillips, S., & Cooney, M. (2005). Aiding peace, abetting violence: Third parties and the management of conflict. *American Sociological Review*, *70*, 334–354.

Rachlin, H. (2015). Social cooperation and self-control. *Managerial and Decision Economics*, n/a-n/a. https://doi.org/10.1002/mde.2714

Roberts, G. (2005). Cooperation through interdependence. *Animal Behaviour*, *70*(4), 901–908.

Tooby, J., Cosmides, L., Sell, A. N., Lieberman, D., & Sznycer, D. (2008). Internal regulatory variables and the design of human motivation: A computational and evolutionary approach. In A. J. Elliott (Ed.), *Handbook of approach and avoidance motivation* (pp. 251–271). Mahwah, NJ: Lawrence Erlbaum Associates.

Supplementary Material

Details on exclusion criteria and decisions

Subjects who did not directly witness the recalled conflict (e.g., conflicts that took place via phone or internet) or who witnessed the conflict at a distance that discouraged or prevented them from intervening (e.g., cases of road rage or overhearing an argument in someone else's house) were excluded. Subjects reporting conflicts involving guns were also excluded because the steep physical costs present in such a situation to deter intervention are likely vastly greater than situations not involving guns, and we wanted to ensure that the effects of WTR and the likelihood of intervening in everyday situations were not washed out by these events, and we didn't have enough cases to be able to generalize to situations involving weapons (but see Phillips & Cooney, 2005). We also excluded subjects who reported conflicts involving multiple transgressors because they introduce complex group dynamics that are beyond the scope of the present paper (i.e., focusing on situations analogous to the third-party punishment game). In addition, we excluded subjects who reported parent/child conflicts (e.g., scolding) due to social norms regarding both parenting styles and interfering with another's parenting, and we excluded conflicts arising during participation in a sport because of the team dynamics involved. Finally, we excluded subjects who reported conflicts in which transgressors and victims had asymmetric institutional power (e.g., boss and employee, professor and student) because of the strong inherent disincentives for subjects to intervene in such situations.

We did not code responses for those subjects that were excluded a priori on the basis of most of the exclusion criteria outlined above. Specifically, the following types of cases were not coded:

- physical barriers that would have made intervention impossible (32)
- conflicts involving a gun (3)
- reported "conflicts" that from the descriptions were actually jokes (4) or not conflicts (12)
- conflicts involving more than one attacker and/or victim (36)
- conflicts that were not witnessed firsthand (23)
- conflicts that were not single events (e.g., "my friend used to get bullied a lot"; 25)
- conflicts between parents/children (7)
- conflict in which a disputant was a police officer (1)
- conflicts in which the subject was either the attacker or the victim (54)
- conflicts that took place during a sports match (19)
- cases where subjects did not respond to the question describing the conflict and/or their own actions, or where it was impossible for the coders to discern from the description what had actually occurred (132)

However, we did code responses for cases involving bosses and teachers/professors before we decided to exclude them (after coding but prior to initial data analyses) due to their inherent power asymmetries

- boss (35)
- teacher/professor (6; but one of these cases [who "did nothing"] did not have a valid wtr_transgressor, so it will be excluded from model containing that term).

The analyses reported in Tables S9 and S10 reinclude these 41 cases. There were 35 cases of doing nothing, 5 interventions, and one punishment. There were no substantive differences between these models and the ones reported in the main text.

Demographic predictors of exclusion

Additionally, we checked whether any of the demographic variables we had available (age, sex, and dataset) predicted meeting exclusion criteria using 3 logistic regression models predicting being excluded.

Sex (dummy coded as 1 = male) did not predict meeting exclusion criteria, $b = .04$, $OR = 1.04$, $p = .750$. Age did predict meeting exclusion criteria, such that older subjects were slightly more likely to be excluded, $b = .03$, $OR = 1.03$, $p < .001$. Finally, dataset also predicted meeting exclusion criteria, with both US students ($b = 1.65$, $OR = 3.91$, $p < .001$) and US MTurkers ($b = 1.36$, $OR = 5.21$, $p < .001$) being significantly more likely to meet exclusion criteria than Japanese students. Additionally, US MTurkers were somewhat more likely to meet exclusion criteria than were US Students, $b = .29$, $OR = 1.33$, $p = .023$.

Table S1. Examples of responses that were coded as “punishment,” “intervention,” or “nothing.”

Code	Conflict	Third Party’s Response
Punish	1. Stranger robbed another stranger	Chased after transgressor
	2. Stranger attempted sexual assault on friend	Yelled at transgressor
	3. Stranger insulted another stranger	Insulted transgressor
	4. Friend cut hair off of stranger	Hit and scolded transgressor
	5. Stranger pushed friend at a club	Fought the transgressor
Intervene	1. Acquaintance drunkenly attacked friend	Broke up confrontation
	2. Acquaintance started verbal argument with friend	Removed friend from situation
	3. Fistfight between acquaintance and friend	Called the police
	4. Argument between acquaintance and friend over a romantic partner	Calmed the situation down by facilitating discussion
	5. Stranger insulting another stranger	Verbally stood up for victim
Nothing	1. Two strangers in a gang-related fight	Took no action
	2. Stranger insulted friend	Consoled friend afterward
	3. Witnessed stranger tackle another stranger in cafeteria fight	Took no action
	4. Witnessed stranger attack a homeless person	Took no action
	5. Witnessed stranger mug stranger in parking lot	Asked if victim was okay afterward

Note. The wording of these reported conflicts and responses has been edited from subject’s exact responses to preserve privacy—subjects were told on the consent form for the experiment that their answers would not be directly quoted in any public dissemination. The edited examples preserve the nature of the situation and response. The majority of responses coded as “nothing” were self-reported by the subject as taking no action. Some cases were coded as “nothing” despite subjects saying they took action, such as the two examples in the table that did not meet our definitions of intervention or punishment for the current paper. “Stranger,” “acquaintance,” and “friend” labels in the table refer to the subject’s relationship with the person involved.

Table S2. Full Ordinal Logistic Regression Model for Intervention and Punishment

Parameter	b	95% CI	OR	p
Intercept 1	-1.96	[-2.43, -1.49]	0.14	< .001
Intercept 2	-5.07	[-5.66, -4.48]	0.01	< .001
WTR_{victim}	1.05	[0.41, 1.68]	2.85	.001
$WTR_{transgressor}$	-0.78	[-3.12, 1.56]	0.46	.514
US MTurkers	0.39	[-0.28, 1.06]	1.48	.256
US Students	1.12	[0.50, 1.74]	3.07	< .001
WTR_{victim}*Transgressor WTR	0.44	[-2.19, 3.06]	1.55	.745
WTR_{victim}*US MTurkers	0.65	[-0.23, 1.53]	1.92	.148
WTR_{victim}*US Students	0.00	[-0.82, 0.82]	0.99	.999
$WTR_{transgressor}$*US MTurkers	1.37	[-1.33, 4.06]	3.92	.320
$WTR_{transgressor}$*US Students	0.62	[-2.11, 3.34]	1.85	.659
WTR_{victim}*$WTR_{transgressor}$*US MTurkers	-1.25	[-4.26, 1.77]	0.29	.418
WTR_{victim}*$WTR_{transgressor}$*US Students	-0.75	[-3.84, 2.34]	0.47	.633

Note. WTR = welfare trade-off ratio. Intercept 1 refers to the log odds of responding with intervention or punishment, relative to responding with doing nothing. Intercept 2 refers to the log odds of responding with punishment, relative to responding with intervention or doing nothing. Reference group = JPN Students.

Table S3. Ordinal Logistic Regression Model for Intervention and Punishment (interactions removed)

Parameter	b	95% CI	OR	p
Intercept 1	-2.09	[-2.41, -1.76]	0.12	< .001
Intercept 2	-5.19	[-5.68, -4.70]	0.01	< .001
WTR_{victim}	1.20	[.90, 1.50]	3.33	< .001
$WTR_{transgressor}$	-0.26	[-.67, .15]	0.77	.211
US MTurkers	0.89	[.54, 1.24]	2.44	< .001
US Students	1.12	[.79, 1.45]	3.07	< .001

Note. WTR = welfare trade-off ratio. Intercept 1 refers to the log odds of responding with intervention or punishment, relative to responding with doing nothing. Intercept 2 refers to the log odds of responding with punishment, relative to responding with intervention or doing nothing. Reference group = JPN Students.

Table S4. Ordinal Logistic Regression for Intervention and Punishment with Conflict Type Added

Parameter	b	95% CI	OR	p
Intercept 1	-2.31	[-2.69, -1.93]	0.10	< .001
Intercept 2	-5.38	[-5.90, -4.86]	0.00	< .001
WTR_{victim}	1.17	[0.87, 1.47]	3.22	< .001
US MTurkers	1.02	[0.61, 1.45]	2.80	< .001
US Students	1.27	[0.86, 1.68]	3.55	< .001
Physical	0.60	[0.04, 1.15]	1.81	.037
US MTurkers*Physical	-0.34	[-1.12, 0.44[	0.71	.390
US Students*Physical	-0.36	[-1.05, 0.33]	0.70	.310

Note. WTR = welfare trade-off ratio. Intercept 1 refers to the log odds of responding with intervention or punishment, relative to responding with doing nothing. Intercept 2 refers to the log odds of responding with punishment, relative to responding with intervention or doing nothing. Physical is a dummy code for conflict type (1 = physical; 0 = verbal). Reference Group = JPN Students

Table S5. Ordinal Logistic Regression for Intervention and Punishment with Social Norm Violation Added (JPN Students and US MTurkers only)

	(1)				(2)			
Parameter	b	95% CI	OR	p	b	95% CI	OR	p
Intercept 1	-2.40	[-2.86, -1.94]	0.09	< .001	-2.36	[-2.77, -1.97]	0.09	< .001
Intercept 2	-6.01	[-6.82, -5.20]	0.00	< .001	-5.98	[-6.76, -5.21]	0.00	< .001
WTR_{victim}	1.23	[0.84, 1.62]	3.43	< .001	1.24	[0.85, 1.62]	3.44	< .001
US MTurkers	0.90	[0.29, 1.50]	2.45	.003	0.84	[0.48, 1.19]	2.31	< .001
Social Norm	0.45	[-0.10, 1.00]	1.56	.113	0.40	[0.02, 0.78]	1.49	.037
US MTurkers*Social Norm	-0.09	[-0.83, 0.65]	0.91	.810	-	-	-	-

Note. WTR = welfare trade-off ratio. Intercept 1 refers to the log odds of responding with intervention or punishment, relative to responding with doing nothing. Intercept 2 refers to the log odds of responding with punishment, relative to responding with intervention or doing nothing. Social Norm is a dummy code for a social norm violation (1 = social norm violated; 0 = social norm not violated). Reference Group = JPN Students. Model 2 dropped the nonsignificant interaction.

Table S6. Full OLS Regression Model for Anger

Parameter	b	95% CI	p
Intercept	2.54	[2.31, 2.77]	<.001
WTR_{victim}	1.31	[0.94, 1.69]	<.001
$WTR_{transgressor}$	-2.59	[-3.60, -1.58]	<.001
US MTurkers	0.05	[-0.32, 0.42]	.791
US Students	-0.58	[-0.94, -0.22]	.002
WTR_{victim}* $WTR_{transgressor}$	0.87	[-0.40, 2.13]	.178
WTR_{victim}*US MTurkers	-0.14	[-0.69, 0.41]	.615
WTR_{victim}*US Students	0.41	[-0.11, 0.94]	.123
$WTR_{transgressor}$*US MTurkers	1.02	[-0.31, 2.36]	.132
$WTR_{transgressor}$*US Students	1.54	[0.15, 2.93]	.030
WTR_{victim} * $WTR_{transgressor}$*US MTurkers	-0.85	[-2.47, 0.77]	.304
WTR_{victim} * $WTR_{transgressor}$*US Students	-0.48	[-2.18, 1.22]	.578

Note. WTR = welfare trade-off ratio. Reference group = JPN Students.

Table S7. OLS Model for Anger with Conflict Type Added

Parameter	b	95% CI	p
Intercept	2.50	[2.27, 2.73]	<.001
WTR_{victim}	1.43	[1.08, 1.77]	<.001
$WTR_{transgressor}$	-2.01	[-2.54, -1.47]	<.001
US MTurkers	0.05	[-0.32, 0.42]	.529
US Students	-0.71	[-1.08, -0.35]	<.001
WTR_{victim}*US MTurkers	-0.26	[-0.76, 0.23]	.302
WTR_{victim}*US Students	0.39	[-0.10, 0.87]	.117
$WTR_{transgressor}$*US MTurkers	0.45	[-0.25, 1.15]	.210
$WTR_{transgressor}$*US Students	1.25	[0.54, 1.97]	.001
Physical	-0.06	[-0.40, 0.27]	.720
Physical*US MTurkers	-0.05	[-0.57, -0.47]	.857
Physical*US Students	0.41	[-0.04, -0.86]	.073

Note. WTR = welfare trade-off ratio. Physical is a dummy code for conflict type (1 = physical; 0 = verbal). Reference Group = JPN Students

Table S8. OLS Model for Anger with Social Norm Violation Added (JPN Students and US MTurkers only)

Parameter	b	95% CI	p
Intercept	2.05	[1.81, 2.29]	<.001
WTR_{victim}	1.07	[0.73, 1.41]	<.001
$WTR_{transgressor}$	-1.53	[-2.06, -1.00]	<.001
US MTurkers	-0.31	[-0.73, 0.10]	.135
WTR_{victim}*US MTurkers	-0.11	[-0.60, 0.37]	.648
$WTR_{transgressor}$*US MTurkers	0.20	[-0.49, 0.88]	.570
Social Norm	1.06	[0.75, 1.36]	<.001
Social Norm*US MTurkers	0.25	[-0.20, 0.70]	.277

Note. WTR = welfare trade-off ratio. Social Norm is a dummy code for a social norm violation (1 = social norm violated; 0 = social norm not violated). Reference Group = JPN Students

Table S9. Ordinal Logistic Regression for Intervention and Punishment with Excluded Subjects Reincluded

Parameter	b	95% CI	OR	p
Intercept 1	-2.17	[-2.48, -1.85]	0.11	< .001
Intercept 2	-5.24	[-5.71, -4.76]	0.01	< .001
WTR_{victim}	1.16	[0.87, 1.45]	3.20	< .001
US MTurkers	0.93	[0.59, 1.27]	2.53	< .001
US Students	1.20	[0.88, 1.52]	3.32	< .001

Note. Model includes 41 additional cases that were excluded from the main text because they were from conflicts containing either a boss or a teacher/professor. Intercept 1 refers to the log odds of responding with intervention or punishment, relative to responding with doing nothing. Intercept 2 refers to the log odds of responding with punishment, relative to responding with intervention or doing nothing. WTR_{victim} (welfare trade-off ratio toward the victim) is a continuous predictor ranging from 0-1.13; sample variables are dummy codes. Reference group = JPN Students

Table S10. OLS Model for Anger with Excluded Subjects Reincluded.

	Reference Group = JPN Students		
Parameter	b	95% CI	p
Intercept	2.51	[2.30, 2.72]	<.001
WTR_{victim}	1.45	[1.12, 1.78]	<.001
US MTurkers	0.05	[-0.28, 0.39]	.749
US Students	-0.59	[-0.92, -0.25]	.001
JPN Students	-	-	-
$WTR_{transgressor}$	-2.12	[-2.63, -1.60]	<.001
WTR_{victim} *US MTurkers	-0.22	[-0.70, 0.26]	.368
Transgressor WTR*US MTurkers	0.57	[-0.11, 1.25]	.103
WTR_{victim} *US Students	0.33	[-0.14, 0.80]	.170
$WTR_{transgressor}$ *US Students	1.35	[0.64, 2.05]	<.001
WTR_{victim} *JPN Students	-	-	-
$WTR_{transgressor}$ *JPN Students	-	-	-

Note. Model includes 40 additional cases that were excluded from the main text because they were from conflicts containing either a boss or a teacher/professor. WTR = welfare trade-off ratio; WTR_{victim} and $WTR_{transgressor}$ are continuous predictors ranging from 0-1.13; sample variables are dummy codes. The two models are identical but recoded with different reference groups.

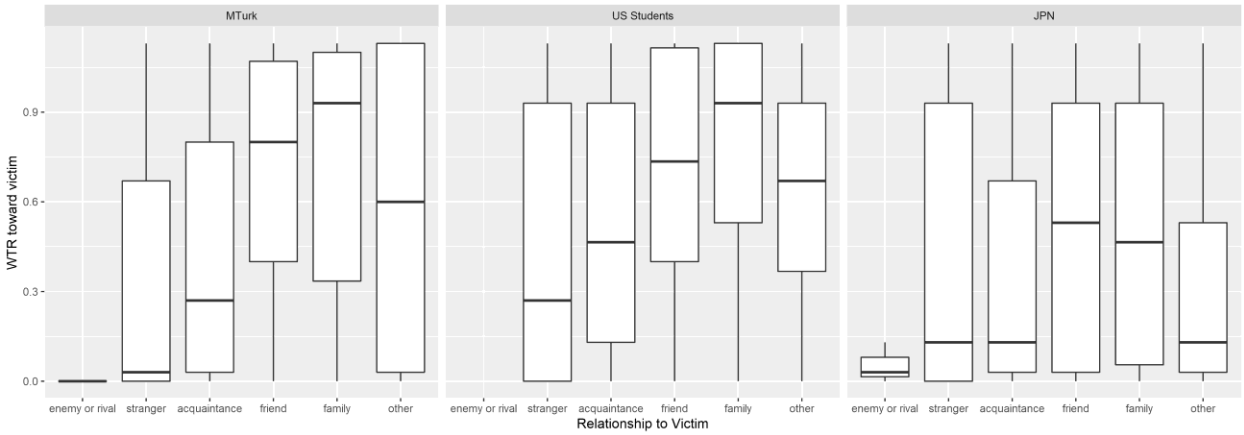

Figure S1. WTR (welfare trade-off ratio) toward the victim as a function of relationship category, broken apart by sample. Boxes represent the inner quartile range (IQR), whiskers extend to the furthest values within 1.5*IQR, and the horizontal lines correspond to the median. Note that US students were not given the option of selecting the “enemy or rival” category.

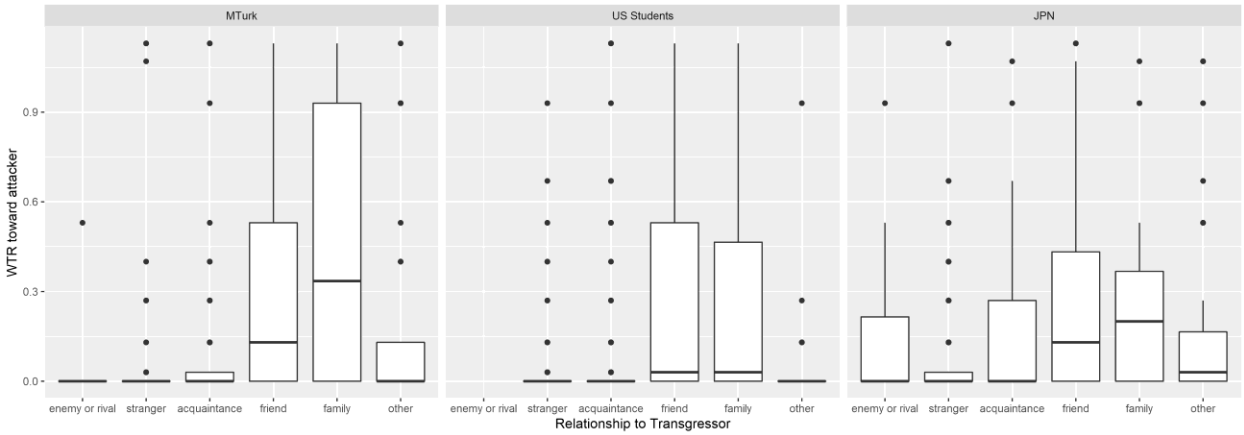

Figure S2. WTR (welfare trade-off ratio) toward the transgressor as a function of relationship category, broken apart by sample. Boxes represent the inner quartile range (IQR), whiskers extend to the furthest values within 1.5*IQR, dots correspond to values beyond 1.5*IQR, and the horizontal lines correspond to the median. Note that US students were not given the option of selecting the “enemy or rival” category.

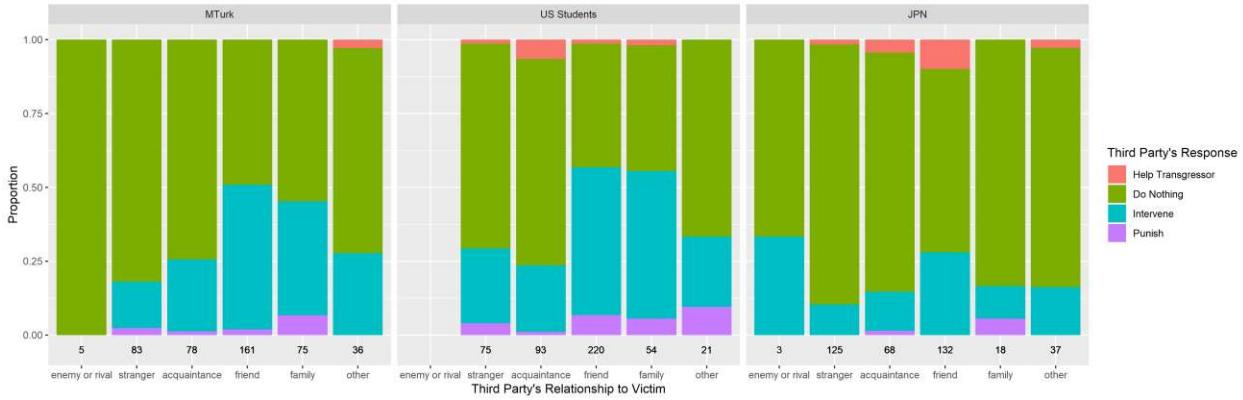

Figure S3. Third party's response as a function of relationship category of victim, broken apart by sample. Colored bars represent the proportion of responses within a category of relationship. Numbers on the x-axis are the cell counts for each category. In all analyses reported in the paper, helping the transgressor was categorized as doing nothing. Note that US students were not given the option of selecting the "enemy or rival" category.

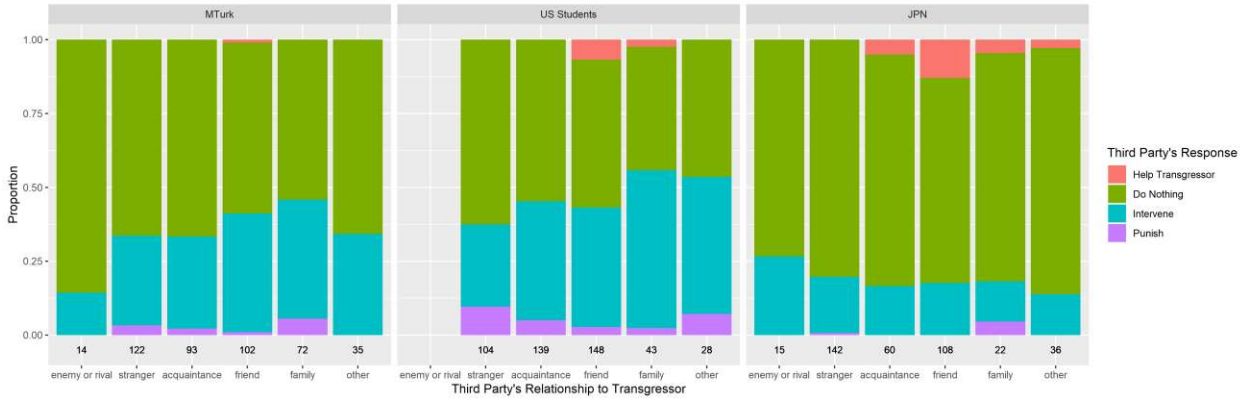

Figure S4. Third party's response as a function of relationship category of transgressor, broken apart by sample. Colored bars represent the proportion of responses within a category of relationship. Numbers on the x-axis are the cell counts for each category. In all analyses reported in the paper, helping the transgressor was categorized as doing nothing. Note that US students were not given the option of selecting the "enemy or rival" category.