Received: 10 July 2022 | Accepted: 21 February 2023

DOI: 10.1111/bjso.12638

the british psychological society
promoting excellence in psychology

ARTICLE

Making sense of punishment: Transgressors' interpretation of punishment motives determines the effects of sanctions

Melissa de Vel-Palumbo[1] | **Mathias Twardawski**[2] | **Mario Gollwitzer**[2]

[1]College of Business, Government and Law, Flinders University, Adelaide, South Australia, Australia

[2]Department of Psychology, Ludwig-Maximilians-Universität, Munich, Germany

Correspondence

Melissa de Vel-Palumbo, College of Business, Government and Law, Flinders University, Adelaide, SA, Australia.
Email: melissa.devel@flinders.edu.au

Funding information

German Federal Ministry of Education and Research; Horizon 2020 Framework Programme, Grant/Award Number: 839639

Abstract

Punishment is expected to have an educative, behaviour-controlling effect on the transgressor. Yet, this effect often remains unattained. Here, we test the hypothesis that transgressors' inferences about punisher *motives* crucially shape transgressors' post-punishment attitudes and behaviour. As such, we give primacy to the social and relational dimensions of punishment in explicating how sanctions affect outcomes. Across four studies using different methodologies (N = 1189), our findings suggest that (a) communicating punishment respectfully increases transgressor perceptions that the punisher is trying to repair the relationship between the transgressor and their group (relationship-oriented motive) and reduces perceptions of harm-oriented and self-serving motives, and that (b) attributing punishment to relationship-oriented (vs. harm/self-oriented, or even victim-oriented) motives increases prosocial attitudes and behaviour. This research consolidates and extends various theoretical perspectives on interactions in justice settings, providing suggestions for how best to deliver sanctions to transgressors.

KEYWORDS

attribution, motives, procedural justice, punishment

BACKGROUND

Punishment is ubiquitous. We discipline children when they misbehave, issue penalties for organizational infractions, and demand even harsher sanctions for criminal offences. From an evolutionary perspective,

punishment functions to enforce social norms and foster cooperative behaviour (van Prooijen, 2017); yet, it does not always have its intended effect. Decades of theorizing and empirical research have largely failed to determine under which conditions punishment 'works' and under which conditions it does not (Sherman, 2017). Unravelling the answer to this question is critical in designing effective sanctioning systems in courts, workplaces, schools, and other institutions.

We propose that attempts to get to the heart of the issue—to understand the variation in reactions to punishment—might benefit from considering how perpetrators subjectively 'make sense' of their punishment. Under this view, punishment likely does not simply function as a contingency-based learning device whereby perpetrators are taught to avoid certain behaviours because it is associated with a negative outcome, as often assumed. Rather, punishment must also be understood as a social interaction in which implicit messages are conveyed to transgressors (Gollwitzer et al., 2011; Sarin et al., 2021; Vidmar & Miller, 1980). Accordingly, the message inferred from punishment will influence how transgressors respond to it (e.g. see Balafoutas et al., 2020).

More specifically, we draw on relational models of punishment to conceptualize inferred punishment purposes in terms of their implications for social relations. We contend that transgressors' attributions to *punisher motives* are an important feature of transgressors' conceptualisation of the purpose of their punishment, and a key driver of their subsequent reactions. Under this framework, inferences about a punisher's intention convey relational information that determines the extent to which punishment can stimulate constructive attitudinal and behavioural change.

Social and relational dimensions of punishment

Punishment can be understood as a social exchange that has relational implications, especially for the transgressor. Wrongdoing threatens transgressors' group membership, making them sensitive to any cues about their social standing (i.e. rejection) that may be implied through punishment. Thereby, the interpersonal dimensions of punishment exchanges are likely to shape the way that transgressors respond to their punishment. For example, social psychologists have shown that addressing threats to one's sense of belonging is crucial in facilitating transgressor openness to reconciliation after wrongdoing (Woodyatt et al., 2017).

The emphasis on social dynamics of justice-related responses is the basis for relational models of procedural and interactional justice (Bies, 2001; Bies & Moag, 1986; Blader & Tyler, 2015). According to these models, satisfaction with authorities and compliance with norms is strongly driven by the treatment of citizens by authorities during decision-making processes (i.e. *how* they punish), more than by the favourability of the outcome itself (Lind & Tyler, 1988). The basic proposition is that the extent to which the decision-making process is perceived as fair determines reactions to decisions (Tyler & Trinkner, 2017).

While procedural justice in a narrow sense refers to the formal and structural properties of a decision-making process (e.g. consistency, bias suppression, accuracy, correctability, representativeness, and ethicality; see Leventhal, 1980), interactional justice refers to how a decision is communicated and whether the interpersonal exchange (e.g. between the judge and a defendant in a court trial) adheres to basic norms of decent conduct (e.g. treating transgressors with respect and dignity). There is an ongoing debate in the literature as to whether procedural and interactional justice are conceptually and empirically separable (see Bobocel & Gosse, 2015; Bobocel & Holmvall, 2001). Here, consistent with our view of punishment as a social exchange, we concentrate on interpersonal treatment (interactional justice)—particularly in our empirical work—while also assuming that insights from any literature that uses a broader conceptualisation of procedural justice (in line with Blader & Tyler, 2003; Tyler & Bies, 2015; Tyler & Wakslak, 2004) hold true in the current context.

Fair treatment is thought to signal to the transgressor that they are valued, promoting their endorsement of the authority as a legitimate representative of the broader social group, as well as identification with that group and, consequently, its norms (Tyler & Blader, 2000). Procedural justice, particularly its interactional elements, has been associated with interpersonal trust following interpersonal transgressions

(Tomlinson, 2012), trust in authorities (Grootelaar & van den Bos, 2018), acceptance of negative outcomes (Greenberg, 1993), positive emotions and organizational loyalty (Chebat & Slusarczyk, 2005), institutional (mis)conduct (Beijersbergen et al., 2015), and lower criminal reoffending (McGrath, 2009).

Motive attributions for punishment

It is already well established that moral judgements of wrongdoing heavily depend on attributions of intent (Cushman, 2008). Here, we argue that being punished triggers a cognitive process by which transgressors look for clues about the intentions or motives a punisher might pursue (Gollwitzer & Okimoto, 2021) and that their response to the punishment will be determined by such attributions—in particular, whether those motive attributions address their relational concerns.

The idea that perceived motives matter in justice-related interactions is not new. In fact, procedural justice theorists explicitly claim, 'the key to people's reactions to authorities lies in their attributions of motives to those authorities' (Tyler, 2003, p. 325). Specifically, fair treatment is thought to influence perceptions of fairness through perceived 'trustworthy' (i.e. benevolent) motives (Tyler, 2008; Tyler & Bies, 2015) that convey messages about a justice recipient's relationship with the authority. In other words, the way a punishment is delivered is influential *because* it contains clues about the authority's intentions.

However, the bulk of procedural justice research is correlational, limiting inferences about causal relationships and psychological mechanisms (Nagin & Telep, 2020). In addition, much of this literature focuses on delivery of unfavourable decisions, rather than sanctions for wrongdoing. The latter is a qualitatively different context, which may attract more defensive and negative attributions, as transgressors grapple with shame, moral condemnation, and threat to belonging. To our knowledge, there have been no empirical demonstrations of the idea that punishment delivered in a just manner has a causal effect on motive attributions, and that these attributions shape transgressors' post-punishment behaviour. And while research has found that fair punishment can lead to a sense of belongingness (van Prooijen et al., 2008), the role of intention attributions in driving these effects has not yet been clearly demonstrated.

We propose that punishment may be attributed to five key punisher motives, with these motives defined in terms of their interpersonal orientation and implications for social relationships. Our approach deviates from other work on punishment purposes (e.g. Carlsmith, 2006) as our model considers the social context of punishment, in particular, transgressors' relational concerns (cf. those which take the perspective of victims or observers of injustice). As per Gollwitzer and Okimoto's (2021) model, which is inspired by work on Social Value Orientation (e.g. Van Lange, 1999), punishment can be attributed to either (1) *relationship-oriented* (prosocial/cooperative), (2) *harm-oriented* (antisocial/competitive), or (3) *self-oriented* (individualistic) motives. Notably, these three attribution categories apply mainly to dyadic situations (in which the victim is the punisher and the transgressor is the punishee). Extending this model to the third-party context, we draw from Oswald et al.'s (2002) empirical work on third-party punishment motives, which identifies motives relating to parties beyond the offender-punisher dyad. Accordingly, as we explain in more detail below, attributions can also be made to (4) *victim-oriented* motives, and (5) *society-oriented* motives. The five motives vary on their implications for transgressors making such attributions, ranging from constructive (e.g. prosocial attitudes towards authorities, adoption of desired norms, and behaviour change) to destructive (e.g. hostility towards authorities, rejection of social norms, antisocial behaviour).

First, punishment can be *relationship-oriented*, reflecting that the punisher is trying to restore positive relations between the transgressor, the victim, and the community to which they belong. Importantly, the key defining feature of this motive—distinguishing it from victim- and society-oriented motives—is the perception that punishment benefits the transgressor. Given that being punished is usually an aversive experience since it entails 'costs' for the transgressor, the idea that punishment may be attributed to prosocial motives at all may sound strange at first glance. However, if enacted appropriately, punishment can convey a message of social inclusion (van Prooijen et al., 2008) and a promise of reintegration into the community once the 'costs' are paid. This also aligns with the concept of reintegrative shaming ('judge

the act, not the person;' Braithwaite, 1989); according to this approach, people will respond favourably to messages of belongingness within punishment that promote a positive relationship with society, and by extension, its norms.

Second, punishment can be *harm-oriented*, meaning that punishment is perceived to be targeted at the transgressor, but primarily with a desire to harm them. This motive maps onto the notion of pure retribution (Carlsmith, 2006): the idea that punishment should close the injustice gap by restoring the 'balance of suffering' (Frijda, 1994, p. 272). Harm-oriented motive attributions are likely those that authorities are typically seeking to avoid through procedurally just treatment—for example, when citizens believe an authority is acting out of 'personal prejudices', driven by animosity rather than acting in the parties' interests (Tyler, 2008, p. 31). Viewing punishment as motivated by harm is likely to be destructive. Disrespectful or degrading treatment can lead to defiance, severing social bonds and resulting in a rejection of authorities and the very values they are trying to promote (Sherman, 1993). In addition, stigma and social rejection leads people to morally disengage from shameful transgressions (Woodyatt & Wenzel, 2013), which may reduce the motivation to change behaviour (Ahmed & Braithwaite, 2004).

Third, punishment can be *self-oriented*, in which the punisher is perceived to be acting primarily in their own self-interest. For example, an authority could be perceived as punishing in order to enhance or protect their authority or responding out of a sense of moral righteousness (Bottoms & Tankebe, 2012). This category captures what have been termed 'insincere' motives for procedurally just treatment, such as an effort to elicit cooperation without genuine concern for the transgressor's wellbeing (Cherney & Murphy, 2011). These sorts of attributions are unlikely to be constructive as they do not offer transgressors genuine restoration and thus do little to energize a shift towards desired norms. And while perceiving a self-oriented motive might not carry the same sting of rejection as harm-oriented punishment, in the third-party context a self-serving punishment could be seen as an abuse of the punisher's position, which might provoke a sense of injustice and hostility. Therefore, attribution of self-oriented motives may have neutral or perhaps even destructive outcomes.

Fourth, punishment can be *victim-oriented*, such that punishers are primarily interested in benefiting victims through punishment. Victim-oriented goals are often key considerations in punishment. For example, transgressors might believe that punishers are seeking to compensate victims for the harm done (Lotz et al., 2011) or to restore victims' equity, power, or status in the group (Okimoto & Wenzel, 2011). Perceiving victim-oriented motives is unlikely to encourage transgressors to change since these motives do not target transgressors' relational concerns. Indeed, framing punishment in compensatory terms does not appear to deter rule violations (Kurz et al., 2014). On the other hand, addressing victim needs could be constructive to the extent that transgressors see value in 'giving back' to victims. However, this may not outweigh offenders' key concerns about their social standing in the broader group.

Lastly, punishment can be *society-oriented*, that is, the punisher is targeting macro-level concerns (such as confirmation of societal values and societal security) rather than micro-level concerns (such as offender re-education or victim security; Orth, 2003; Vidmar & Miller, 1980). Such attributions could include judgements that an authority is using punishment as a way to 'repay' society for the harm done, to maintain social order by deterring potential wrongdoing in the future (i.e. general deterrence), or to reaffirm group values (Okimoto & Wenzel, 2009). Transgressors making these types of attributions might respond positively to the authority, judging that by making amends, others might forgive them and allow them to re-join the group. In line with this, some research suggests that transgressor are more likely to endorse punishment when it is interpreted as an opportunity to do something for the community (Griffin, 2006). Arguably, however, a society-serving motive might not accomplish much without some implication of the offender's reintegration at the individual level. Therefore, the effect of attributed society-oriented motives on outcomes is also unclear a priori.

Overview of the research

In the present research, we investigate how the way punishment is delivered influences attributions transgressors make about why they are being punished, and how such attributions influence their

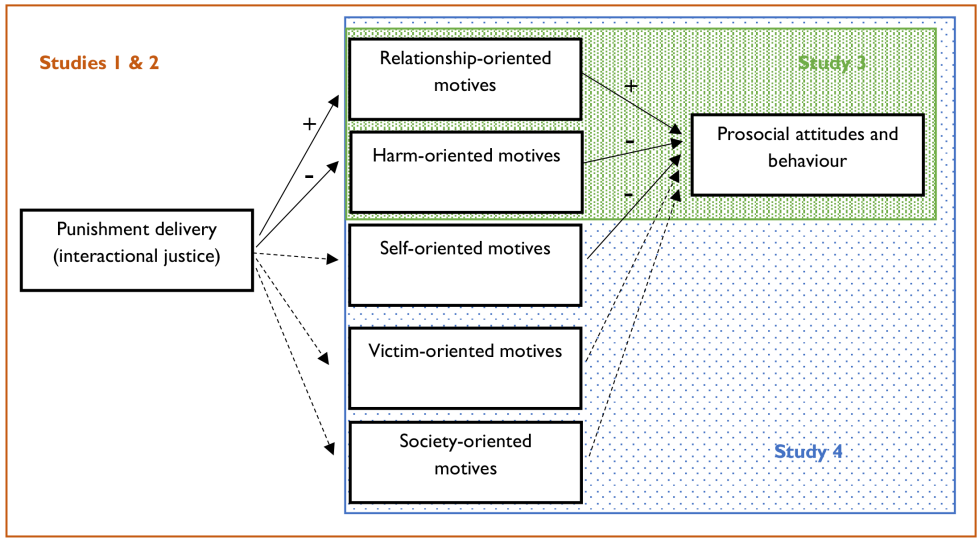

FIGURE 1 Conceptual model: Punishment delivery influences prosocial attitudes and behaviour through motives transgressors attribute to their punishment. Dotted arrows indicate possible effects but without sufficient evidence with which to make clear predictions. In Studies 1 and 2, we test the entire model, while in Studies 3 and 4, we focus on testing the second half (the links between motive attributions and outcomes).

post-punishment attitudes and behaviour. In doing so, our work highlights motive attributions as a key dimension of how transgressors make sense of and react to their punishment. Our model stipulates that (1) the way that punishment is delivered (e.g. communicated by the punisher) affects the extent to which this punishment is attributed to each of the five motives discussed above, and that (2) motive attributions predict transgressors' attitudes and behaviour.

More precisely, we posit, firstly, that delivering punishment in line with procedural (or more specifically, interactional) justice principles is particularly likely to make transgressors attribute the punishment to relationship-oriented motives and, by contrast, least likely to harm-oriented motives. As discussed, theoretical arguments that would allow us to make strong predictions about an effect of just versus unjust punishment delivery on self-, victim-, or society-oriented motives are harder to derive from the literature. Therefore, in the present paper, we focus on hypothesising an effect of just versus unjust delivery of punishment on relationship- and on harm-oriented motive attributions, while exploring the effects on the remaining three attributions.

The second part of our model assumes that each of the five motives to which transgressors may attribute the punishment predicts transgressors' prosocial attitudes towards the punishment (operationalized herein as acceptance of the punishment and of the authority) and behaviour (operationalized as behavioural change or intention to change one's behaviour). More precisely, attributing punishment to relationship-oriented motives should have the strongest (positive) effect on attitudes and (intended) behavioural change, while harm- and, to a lesser extent, self-oriented attributions, should have a negative effect. The effects of victim- and society-oriented attributions are less clear, so we did not make predictions in either direction. Figure 1 displays our theorizing as a path model.

We tested our model using methodological triangulation across four studies. First, we conducted a cross-sectional survey of people who had recently been punished by the criminal justice system. This study established relationships between key constructs as specified in our model at a correlational level. While this methodological design does not permit causal inferences, it has strong ecological validity.

Studies 2, 3 and 4 follow an *experimental causal-chain design* (Spencer et al., 2005) to demonstrate the psychological process underlying procedural justice theory by manipulating the independent variable and the mediator in separate steps. Study 2 used a vignette design to demonstrate that punishment delivery (i.e.

communicating punishment in line with interactional justice criteria) causally affects motive attributions. We also tested the indirect effects of punishment delivery on attitudes and behavioural intentions via transgressors' attributions of the punishment. In Studies 3 and 4, we manipulated motive attributions and observed subsequent causal effects on perpetrator attitudes and prosocial behaviour. We first focused on the effects of relationship and harm-oriented motives in Study 3 (using a hypothetical vignette design), while in Study 4 we explored associations between other motive attributions and outcomes (using a behavioural economic game paradigm).

All measures, manipulations, and exclusions in the studies are disclosed in this article. All studies were pre-registered; all deviations from the plans are disclosed (two in Study 4).

STUDY 1

In Study 1 we tested our full theoretic model (see Figure 1) using a cross-sectional design. We collected data from a sample of people who had recently been sanctioned during *face-to-face interactions* with a court official or police—this ensured social dynamics (and thus interactional justice factors in particular) were present during punishment administration. We measured perceived interactional justice, and participants' attributions of the punishment to either relationship-, harm-, self-, victim-, and society-oriented motives. In addition, we measured three outcomes: acceptance of the authority/sanction (i.e. a measure of prosocial attitudes), and motivation to change (i.e. a measure of prosocial behavioural intention). We also measured actual antisocial behaviour since the punishment—reflecting a rejection of social norms (which we expected to be negatively associated with relationship-oriented motives, and positively associated with harm-oriented motives).

As per our conceptual model, we focused on relationship- and harm-oriented attributions as key attributions relevant to interactional justice, while also exploring the associations between the other three attributions and outcomes. The study pre-registration is available at https://aspredicted.org/im68t.pdf.

Method

Participants

First, we invited people from the US, UK, Canada, Australia, and New Zealand using two online platforms (Mechanical Turk and Prolific Academic) to participate in an initial short online questionnaire for screening purposes. Of all respondents to this survey (n = 3728), only 9% (n = 322) were eligible (i.e. had been recently sanctioned) and were, hence, invited to the main study. From these invited participants, 233 completed the main study (72% response rate), but 15 respondents failed our pre-registered attention check, leaving an N of 218 in our final sample (M_{age} = 34.81, SD = 10.56; 60% female, 40% male; 81% US, 12% UK, 2% Canada, 5% Australia, 1% New Zealand). Most offences for which participants had been adjudicated involved traffic and vehicle violations (42%); the rest were violent (5%); drug (5%); public order offences (4%); dangerous/negligent acts (2%); theft (1%); other categories each represented less than 1% of the sample; offence not specified (38%). Participants were compensated USD$1.10 for the study.

Our sample size provided 53% power to detect significant indirect effects of interactional justice on outcomes via relationship- and harm-oriented motives in a parallel mediation model, assuming correlations of 0.3 between the constructs and 99% confidence intervals (Schoemann et al., 2017). According to G*Power, statistical power to detect bivariate correlations of this magnitude was high (>99%; Faul et al., 2007).

Materials

Full versions of all scales are provided in the Supplementary Materials. Note that all scale items across all four studies, except for antisocial behaviour in Study 1, were measured on 5-point Likert scales ranging from 'Strongly disagree' to 'Strongly agree'.

Interactional justice

We used the 4-item interpersonal justice scale from Colquitt (2001) to measure the interactional aspects of procedural justice (α = .92). Example item: 'He/she spoke to me with respect.'

Punishment motive attributions

Five scales were created to measure perceived relationship-, harm-, self-, victim-, and society-oriented motives for punishment, with reference to the theoretical constructs underlying these attributions. Each scale contained four items. Example items: 'My manager reacted like this in order to… put things right between me and the community' (relationship); '…humiliate me' (harm); '…demonstrate his/her power' (self); '…stand up for victims of crime' (victim); '…maintain order in society' (society).

A confirmatory factor analysis indicated support for the hypothesised five-factor structure, with items loading onto factors representing relationship- (α = .90), harm- (α = .91), self- (α = .88), victim- (α = .88), and society-oriented motives (α = .77). To improve model fit, we excluded one poorly fitting item from the society-oriented scale. Noting strong correlations between some of the motive scales (see Table 1), we also tested several alternate measurement models, including a 4-factor model in which harm- and self-oriented motives loaded onto the same factor, and another in which relationship- and society-oriented motives loaded on the same factor. These models were significantly inferior to the 5-factor model. For full details of all factor analyses see Appendix C in the Supplementary Materials.

Acceptance

Two items were used to measure acceptance of the authority and the sanction (r = .52, p < .001), for example: 'I willingly accepted the decision made.'

Motivation to change

Drawing from the clinical psychology literature on behavioural change (Prochaska et al., 2015), we developed an eight-item scale designed to reflect readiness for and commitment to prosocial behaviour (α = .90). Items tapped into intended compliance with the violated rule, commitment to organizational norms more broadly, accepting responsibility, and seeing punishment as an opportunity to change. Example item: 'I am committed to doing the right thing from now on.'

Antisocial behaviour

Participants were asked how often they had engaged in six antisocial behaviours (e.g. 'used marijuana or some other drug') since the punishment was delivered (Reisig et al., 2014; α = .84). Frequency was reported using a 4-point response scale ranging from 'Never' to 'Often'.

Data integrity measures

Several data integrity measures were included. Participants were unable to participate if they did not affirm in writing that they would read all text carefully (Zhou & Fishbach, 2016) or failed an instructed response item embedded in the questionnaire (Meade & Craig, 2012).

Results

First, examining the link between interactional justice and motive attributions, our hypothesis was supported. Interactional justice was positively correlated with relationship-oriented motives and negatively correlated with harm-oriented motives (see Table 1).

Turning to the link between attributed motives and outcomes, findings generally supported our predictions. Relationship-oriented motives were positively correlated with acceptance and motivation to change. Harm-oriented motives were negatively correlated with acceptance, motivation to change, and positively

predicted antisocial behaviour. However, going against predictions, relationship-oriented motives were positively associated with antisocial behaviour, though this was a weak correlation.

We used the PROCESS macro for SPSS (Hayes, 2013; model 4), with 10,000 bootstrapped re-samples and 99% bias-corrected confidence intervals to test for predicted mediation models (see Table 2). For acceptance and antisocial behaviour, only the unique indirect effect via harm-oriented motives was significant. For motivation to change, only the unique indirect effect via relationship-oriented motives was significant. Overall, therefore, interactional justice conveyed messages that the punishment was aimed at repairing the relationship between the offender and others, while reducing perceptions that the punishment was aimed at harming the transgressor, and these attributions had implications for prosocial attitudes and behaviour.

TABLE 1 Scale correlations (Study 1).

	r							
Variables	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1 Interactional justice	.55**	−.45**	−.46**	.33**	.59**	.36**	.44**	−.02
Motive attributions								
2 Relationship	1	−.13	−.21*	.63**	.69**	.26**	.56**	.15*
3 Harm		1	.76**	.27**	−.24**	−.43**	−.15*	.49**
4 Self			1	.09	−.26**	−.40**	−.23**	.38**
5 Victim				1	.54**	.01	.44**	.37**
6 Society					1	.38**	.50**	.05
7 Acceptance						1	.27**	−.19*
8 Motivation to change							1	−.04
9 Antisocial behaviour								1

*$p < .01$.
**$p < .001$.

TABLE 2 Mediation model statistics: effects of punishment delivery on outcomes via motive attributions (Study 1).

	***M*→dv**			**IV→DV (total effect)**			**IV→DV (direct effect)**			**Indirect effect**		
Variables	***B***	***SE***	**$CI_{99\%}$**	***B***	***SE***	**$CI_{99\%}$**	***B***	***SE***	**$CI_{99\%}$**	***B***	***SE***	**$CI_{99\%}$**
Acceptance	–	–	–	0.38	0.07	0.20, 0.55	0.12	0.09	0.10, 0.34	–	–	–
Relationship	0.15	0.07	−0.04, 0.34	–	–	–	–	–	–	0.09	0.05	−0.05, 0.22
Harm	−0.34	0.07	−0.51, −0.17	–	–	–	–	–	–	0.17	0.04	0.08, 0.27
Motivation to change	–	–	–	0.36	0.05	0.23, 0.49	0.15	0.06	0.01, 0.31	–	–	–
Relationship	0.36	0.05	0.22, 0.49	–	–	–	–	–	–	0.21	0.04	0.11, 0.33
Harm	−0.01	0.05	−0.13, 0.12	–	–	–	–	–	–	0.01	0.02	−0.05, 0.06
Antisocial behaviour	–	–	–	−0.01	0.04	−0.13, 0.10	0.12	0.05	−0.01, 0.25	–	–	–
Relationship	0.08	0.04	−0.03, 0.19	–	–	–	–	–	–	0.05	0.02	−0.01, 0.12
Harm	0.35	0.04	0.25, 0.45	–	–	–	–	–	–	−0.17	0.03	−0.25, −0.10

Note: IV→*M* for relationship- ($B = 0.59$, $SE = 0.06$, $CI_{99\%} = 0.43, 0.74$) and harm-oriented motives ($B = -0.49$, $SE = 0.07$, $CI_{99\%} = -0.67, -0.32$).

Secondary analyses

We also explored relationships between other motive attributions and outcomes in a multivariate analysis, regressing outcomes on all five attribution scales (see Table 3). Potentially due to the strong correlation between self- and harm-oriented motive scales, self-oriented motives did not predict any outcome. Society-oriented motives were seemingly constructive (though this was only evident for acceptance; marginal for motivation to change), while results for victim-oriented motives were mixed, with both constructive (on motivation to change) and destructive implications (on antisocial behaviour).

STUDY 2

Study 1 confirmed that interactional justice was positively correlated with relationship-oriented motives and negatively correlated with harm-oriented motives, and these attributions had implications for prosocial attitudes and behaviour. Study 2 tested whether the way punishment is delivered (in this case, the way it is communicated) causally influences motive attributions, attitudes, and (intended) behaviour. We also manipulated the *severity* of the punishment in order to assess the impact of interpersonal dimensions of the sanctioning process relative to mere outcomes (prior research has indicated that procedural justice effects are stronger for more unfavourable outcomes; Brockner & Wiesenfeld, 1996). The study thus consisted of a 2 × 2 (just vs. unjust punishment delivery style × low vs. high severity punishment) between-groups design with random allocation to conditions. We focused on effects on (and indirect pathways through) relationship- and harm-oriented motives, though we also explored the role of other motives. The study used vignettes involving hypothetical punishment for an organizational infraction. From that perspective, we measured punishment motive attributions, acceptance of the authority and the sanction, and motivation to change. This study was pre-registered at https://aspredicted.org/yg2fh.pdf.

Method

Participants

We recruited UK residents from Prolific Academic. Eligible participants were compensated £0.70 for their participation. After excluding 26 participants who failed pre-registered exclusion criteria, the final sample size was 474 (M_{age} = 34.66; *SD* = 11.69; 68% female, 32% male; <1% other).

The effect requiring the most statistical power is the interaction effect between the two experimental factors, though this effect is not central to our model. We had predicted an attenuated effect; that is, that there would be positive effects of communication at both levels of punishment severity, but the effect would be stronger at high levels of severity. To conduct a power analysis for this analysis, we followed Perugini et al. (2018). We assumed that the larger of the conditional effects would be equivalent in size to an *f* of 0.39–0.49 (based on correlations between interactional justice and outcome variables in Study 1).

TABLE 3 Multivariate regression analysis (Study 1).

	Acceptance			Motivation to change			Antisocial behaviour		
Predictor	**β**	***SE***	***p***	**β**	***SE***	***p***	**β**	***SE***	***p***
Relationship	.06	0.09	.498	.31	0.07	<.001	.10	0.06	.269
Harm	−.17	0.10	.079	−.02	0.07	.833	.37	0.06	<.001
Self	−.15	0.09	.092	−.13	0.07	.152	.10	0.05	.263
Victim	−.15	0.09	.113	.18	0.07	.038	.20	0.06	.024
Society	.37	0.10	<.001	.16	0.07	.055	−.01	0.06	.945

Given a 50% attenuated interaction is approximately four times smaller than the larger effect; therefore, according to G*Power (Faul et al., 2007), our sample had 58%–74% power to detect at least 50% attenuation (i.e. $f = 0.10$–0.12 for an ANOVA model with $\alpha = .05$).

The more critical test of our model in this study required a significant conditional effect of communication in both severity conditions. A sensitivity power analysis using G*Power indicated indicated that our sample had 80% power to detect conditional effects as small as $d = 0.23$ (one-tailed t-test for independent means with $\alpha = .05$).

Materials

Vignettes

We created hypothetical vignettes in which participants were asked to imagine they stole money from their co-workers during a work shift at a café. During a 'disciplinary meeting', during which a punishment is issued, the manager communicates with them in either (a) a just manner, or (b) a unjust manner, constituting the punishment communication manipulation. The outcome was described as either (a) low in severity (the protagonist will not receive tips on their next shift), or (b) high in severity (the protagonist will not receive tips for the whole next month), constituting the punishment severity manipulation. The communication manipulation text was informed by the interactional justice literature, incorporating notions of respect and propriety in communication (i.e. the interpersonal aspects of procedural justice). For example, in the fair condition, the manager's tone is polite (vs. rude), perspective-taking is expressed (vs. not), the act is described in less morally loaded terms ('keeping tips' vs. 'stealing from other employees'), and there is an absence of disrespectful remarks ('I'm not going to fire you for this' vs. 'You're lucky I don't fire you right now!').

Punishment motive attributions

Attribution scale items were amended to fit the organizational context of the study. Only four of the 20 items were substantially different to those in Study 1.

A confirmatory factor analysis again indicated support for the hypothesised 5-factor structure, with items loading onto factors representing relationship- ($\alpha = .75$), harm- ($\alpha = .86$), self- ($\alpha = .76$), victim- ($\alpha = .66$), and society-oriented motives ($\alpha = .62$). We excluded two poorly fitting items: one from the victim- and one from the society-oriented scale. Full details of the factor analysis are presented in Appendix C in the Supplementary Materials.

Acceptance

A six-item scale was used to measure acceptance of the authority and the sanction ($\alpha = .94$). Three of these items were taken from Tyler and Wakslak (2004), while the remaining three were added to increase psychometric properties of the scale. Example items: 'I have no hard feelings towards my manager'; 'The outcome I received was fair'.

Motivation to change

We used the same eight-item scale as in Study 1, with some wording changes to reflect the new context ($\alpha = .88$).

Manipulation checks

Three items from Study 1 (adapted from Colquitt, 2001) were used as a manipulation check for interactional justice ($\alpha = .95$), and we also designed a three-item scale to check the outcome severity manipulation ($\alpha = .80$).

Data integrity measures

As planned, participants were excluded if they: did not affirm in writing that they would read all text carefully (Zhou & Fishbach, 2016); failed a simple reading check (placed after the vignette) four times or more; or failed an instructed response item (Meade & Craig, 2012).

Results

Interactional justice was rated as higher in the just communication condition ($M = 4.57$, $SD = 0.56$) than in the unjust condition ($M = 2.69$, $SD = 0.97$), $t(374.94) = -25.74$, $p < .001$, $d = 2.37$. The outcome was rated more severe in the high severity condition ($M = 3.10$, $SD = 0.98$) than in the low severity condition ($M = 2.57$, $SD = 0.98$), $t(471.96) = -5.91$, $p < .001$, $d = 0.54$. There were no interactive effects of the manipulations on the manipulation checks. Correlations between key measures are presented in Table 4.

Main effects of punishment communication and severity

General linear models were conducted to produce estimates of main effects for the two manipulations and their interaction term. Estimated marginal means are reported in Table 5. In line with our hypotheses, interactionally just punishment increased attributions to relationship-oriented motives, $F(1, 470) = 89.24$, $p < .001$, $\eta_p^2 = 0.16$, and reduced attributions to harm-oriented motives $F(1, 470) = 63.64$, $p < .001$, $\eta_p^2 = 0.12$.

There was no main effect of communication on victim-oriented motives, $F(1, 470) = 0.36$, $p = .550$, $\eta_p^2 = 0.001$. Interactionally just communication was also perceived as less self-serving, $F(1, 470) = 37.68$, $p < .001$, $\eta_p^2 = 0.07$, while increasing perceptions that punishment was society-oriented, $F(1, 470) = 9.29$, $p = .002$, $\eta_p^2 = 0.02$; however, these were far smaller in magnitude than the effects on relationship- and harm-oriented motives. Thus, as hypothesised, interactional justice primarily influenced perceptions that punishment is targeted at the transgressor, either in a benevolent (relationship-oriented) way, or a malevolent (harm-oriented) way. Punishment severity did not moderate any of these effects nor did it have any main effects ($ps \geq .05$).

As hypothesised, interactionally just punishment increased acceptance, $F(1, 470) = 38.46$, $p < .001$, $\eta_p^2 = 0.08$, and motivation to change, $F(1, 470) = 11.15$, $p < .001$, $\eta_p^2 = 0.02$. Severity reduced acceptance, $F(1, 470) = 9.13$, $p = .003$, $\eta_p^2 = 0.02$, but did not influence motivation to change, $F(1, 470) = 0.60$, $p = .438$, $\eta_p^2 = 0.001$. The interaction between communication and severity was nonsignificant for both acceptance, $F(1, 470) = 0.02$, $p = .877$, $\eta_p^2 < 0.001$, and motivation to change, $F(1, 470) = 0.17$, $p = .682$,

TABLE 4 Scale correlations (Study 2).

	r					
Variables	**2**	**3**	**4**	**5**	**6**	**7**
Motive attributions						
1 Relationship	−.46**	−.30**	.40**	.55**	.58**	.51**
2 Harm	1	.70**	−.12*	−.37**	−.58**	−.40**
3 Self		1	−.06	−.19**	−.47**	−.34**
4 Victim			1	.44**	.29**	.25**
5 Society				1	.51**	.45**
6 Acceptance					1	.60**
7 Motivation to change						1

*$p < .01$.

**$p < .001$.

TABLE 5 Estimated marginal means for main effects of punishment manipulations (Study 2).

	EMM (*SE*)			
	Punishment communication		Punishment severity	
Dependent variable	Unjust (*n* = 236)	Just (*n* = 238)	Low (*n* = 236)	High (*n* = 238)
Motive attributions				
Relationship	3.40 (0.05)	4.02 (0.05)	3.73 (0.05)	3.69 (0.05)
Harm	2.54 (0.06)	1.91 (0.06)	2.17 (0.06)	2.28 (0.06)
Self	2.90 (0.05)	2.45 (0.05)	2.68 (0.05)	2.67 (0.05)
Victim	3.69 (0.05)	3.73 (0.05)	3.72 (0.05)	3.70 (0.05)
Society	3.99 (0.04)	4.17 (0.04)	4.10 (0.04)	4.05 (0.04)
Acceptance	3.81 (0.05)	4.28 (0.05)	4.16 (0.05)	3.93 (0.05)
Motivation to change	4.14 (0.04)	4.34 (0.04)	4.26 (0.04)	4.22 (0.04)

Note: Models include the communication × severity interaction term.

$\eta_p^2 < 0.001$. Conditional effects of communication were small to moderate across both severity conditions (*d*s for acceptance = −0.60 [low severity], −0.54 [high severity]; *d*s for motivation to change = −0.35 [low severity], −0.27 [high severity]). This suggests that interactional justice effects are generalisable across the severity levels tested in this study; *how* punishment was delivered, not the mere outcome, was critical in determining transgressor reactions to punishment.

Indirect effects of punishment communication on outcomes via motive attributions

We used the PROCESS macro for SPSS (Hayes, 2013; model 4 for parallel mediators with 10,000 bootstraps and 99% bias-corrected confidence intervals) to test which motive attributions could account for the effects of punishment communication on acceptance and motivation to change (see Table 6). Results were generally consistent with our theorizing—interactionally just punishment had constructive outcomes, and these were explained by increased attributions of the punishment to relationship-oriented motives (for both acceptance and motivation to change) and reduced attributions to harm-oriented motives (for acceptance only). There were also (smaller) indirect effects of interactional justice on acceptance and motivation to change via a decrease in self-oriented motives, as well as even smaller yet still-significant indirect effects via an increase in society-oriented motives. The indirect effects via victim-oriented motives were not significant, and nor did victim-oriented motives appear to predict outcomes. Combined, the indirect effects fully mediated the total effects of the communication manipulation on both outcomes.

STUDY 3

While Studies 1 and 2 provided evidence for the proposition that (perceived) punishment motives influence perpetrators' reactions to punishment, (attributed) punishment motives were not manipulated. Thus, Study 3 employed an experimental design that allowed for a test of the causal effects of motive attributions on perpetrator attitudes and behaviour.

Based on existing literature and findings so far, we judged that manipulating perceptions of relationship- versus harm-oriented punishment would create a meaningful contrast. The study thus consisted of a 2-cell (relationship- vs. harm-oriented punishment) between-groups design. Participants read vignettes describing a hypothetical punishment for an institutional rule violation (cheating on a university exam). We assessed two outcomes: acceptance of the authority and the sanction, and motivation to change. This study was pre-registered at https://aspredicted.org/yq8ju.pdf.

TABLE 6 Mediation model statistics: effects of punishment communication on outcomes via motive attributions (Study 2).

	M→dv			IV→DV (total effect)			IV→DV (direct effect)			Indirect effect		
Variables	***B***	***SE***	**$CI_{99\%}$**	***B***	***SE***	**$CI_{99\%}$**	***B***	***SE***	**$CI_{99\%}$**	***B***	***SE***	**$CI_{99\%}$**
Acceptance	–	–	–	0.48	0.08	0.28, 0.68	−0.01	0.06	−0.17, 0.16	–	–	–
Relationship	0.32	0.05	0.19, 0.46	–	–	–	–	–	–	0.20	0.04	0.11, 0.31
Harm	−0.23	0.05	−0.36, −0.11	–	–	–	–	–	–	0.15	0.04	0.05, 0.26
Self	−0.18	0.05	−0.30, −0.05	–	–	–	–	–	–	0.08	0.03	0.02, 0.16
Victim	0.05	0.04	−0.06, 0.15	–	–	–	–	–	–	0.002	0.01	−0.01, 0.02
Society	0.27	0.06	0.13, 0.42	–	–	–	–	–	–	0.05	0.02	0.01, 0.11
Motivation to change	–	–	–	0.19	0.06	0.04, 0.34	−0.11	0.05	−0.25, 0.03	–	–	–
Relationship	0.27	0.04	0.17, 0.38	–	–	–	–	–	–	0.17	0.03	0.10, 0.26
Harm	−0.05	0.04	−0.15, 0.05	–	–	–	–	–	–	0.03	0.03	−0.04, 0.10
Self	−0.13	0.04	−0.23, −0.03	–	–	–	–	–	–	0.06	0.02	0.01, 0.12
Victim	0.003	0.03	−0.09, 0.09	–	–	–	–	–	–	0.001	0.003	−0.01, 0.01
Society	0.21	0.05	0.09, 0.34	–	–	–	–	–	–	0.04	0.02	0.01, 0.09

Note: IV→*M* for relationship- ($B = 0.62$, $SE = 0.07$, $CI_{99\%} = 0.45, 0.79$), harm- ($B = -0.63$, $SE = 0.08$, $CI_{99\%} = -0.84, -0.43$), self- ($B = -0.45$, $SE = 0.07$, $CI_{99\%} = -0.65, -0.26$), victim- ($B = -0.04$, $SE = 0.07$, $CI_{99\%} = -0.15, 0.24$), and society-oriented motives ($B = 0.18$, $SE = 0.06$, $CI_{99\%} = 0.03, 0.33$). IV coded unjust = 0, just = 1.

Method

Participants

We recruited participants from the US, UK, Canada, Australia, and New Zealand using Prolific Academic. Eligible participants were compensated £0.70 for their participation. After excluding nine participants who failed pre-registered exclusion criteria, the final sample size was 285 ($M_{age} = 40.32$; $SD = 12.94$; 51% female, 49% male; <1% other).

A sensitivity power analysis using G*Power (Faul et al., 2007) indicated that our sample provided 80% power to detect an effect size of at least $d = 0.30$ for a one-tailed difference test between two independent groups with an alpha level of .05. This is comparable to the smallest relationship between motives (relationship- and harm-oriented) and outcomes (acceptance and behavioural intention) in Study 1 (smallest $r = .15$, equivalent to $d = 0.30$) and thus a reasonable sample to detect the types of effects we expected.

Materials

Vignettes

We created hypothetical vignettes in which participants were asked to imagine they are caught cheating on a university exam. The punishment—a fail for the course—is issued by the course instructor. Participants

were informed that the course instructor either has (a) a positive reputation among students, and acts with seemingly relationship-oriented motives, or (b) a negative reputation among students, and acts with seemingly harm-oriented motives.

Outcome measures

Acceptance (α = .87) and motivation to change (α = .83) were measured using the same scales as in Study 2, with minor tweaks to match the different context. The two measures were moderately correlated $r = .39$, $p < .001$. We did not measure punishment motive attributions since our manipulations showed strong validity in pre-testing (see Appendix C in the Supplementary Materials).[1]

Data integrity measures

As planned, participants were excluded if they: did not affirm in writing that they would read all text carefully (Zhou & Fishbach, 2016); failed at least one of two simple reading checks (placed after the vignette); failed an instructed response item (Meade & Craig, 2012); or failed a simple English proficiency test.

Results

In line with our predictions, a one-tailed *t*-test indicated that participants in the relationship-oriented condition accepted their punishment more than those in the harm-oriented condition (see Table 7). Those in the relationship-oriented condition also showed stronger motivation to change on average, but this effect did not reach statistical significance. The study was underpowered to detect an effect of this magnitude.

STUDY 4

In the final study, we aimed to replicate the results of Study 3 in a new context with lower demand characteristics and including a behavioural measure. Again, we sought to contrast relationship- versus harm-oriented punishment. However, we also took into account the overlap between self- and harm-oriented motives found in the first three studies. We, therefore, ultimately used a punishment manipulation that, while on face value is best described as a self-oriented punishment, was attributed to both harm- and self-oriented motives according to pre-testing (see Appendix C in the Supplementary Materials for a detailed description of the pre-test). We also assessed participants' attributions of the punisher's motive to punish since there was some potential ambiguity in our manipulations.

Participants played a public goods game representing a social dilemma between behaving cooperatively (i.e. maximizing the joint payoff of all players by contributing one's resources to a common resource pot—the 'public good'—which is divided equally among all players) or egoistically (i.e. maximizing one's individual payoff by keeping one's resources for oneself, but profiting from the public good nonetheless;

TABLE 7 Between-group statistics (Study 3).

	***M* (*SD*)**					
Dependent variable	**Relationship-oriented (*n* = 143)**	**Harm-oriented (*n* = 142)**	***t***	***df***	***p***	***d***
Acceptance	4.47 (0.68)	3.97 (0.71)	−5.98	283	<.001	0.71
Motivation to change	4.30 (0.61)	4.22 (0.55)	−1.16	283	.123	0.14

[1]According to pre-test ($N = 197$) results, the manipulations differed on relationship-, harm-, and self-oriented motives only. In addition, factor analysis of motive attribution items in the pre-test indicated support for the hypothesised 5-factor model. The 5-factor model performed significantly better than alternate models (see Supplementary Materials).

see Komorita & Parks, 1994). Egoistic choices by one player are considered unfair and often punished by other players, even at the expense of their own resources (e.g. Brandt et al., 2006).

Punishment was delivered to participants acting egoistically by a 'referee', under one of two different sanctioning systems (randomly assigned), which we refer to as *self-profit* versus *group-profit* conditions. In both systems, points are deducted from the egoistic player as punishment for their behaviour, but the motive differs from the perspective of the transgressor (the participant). Under the group-profit sanctioning system, the deducted points are then redistributed equally among all players in the team (including the participant), reflecting the notion of punishment as relationship-oriented. Under the self-profit sanctioning system, deducted points are kept by the referee, reflecting the notion of punishment as harm−/self-oriented. We pre-tested these manipulations (see Appendix C in the Supplementary Materials), which also allowed us to refine the experimental design, for example by ensuring the game instructions and structure were clearly understood by participants before conducting the main experiment.

Our dependent variables were participants' acceptance of the punishment, and prosocial behaviour (defined here as participants' contribution to the public good in a subsequent round of the game). Our main hypothesis was that acceptance and prosocial behaviour would be higher in the group-profit than in the self-profit sanctioning system. The study was pre-registered at https://aspredicted.org/AKL_VC58xr2.pdf.

Method

Participants

We ran 248 individuals through the study—22 of whom were ineligible for punishment since they contributed their full endowment to the shared pot. Twenty-seven further participants were excluded using our pre-registered exclusion criteria (see below), leaving 199 participants with a mean age of (SD = 8.99; 43% female, 57% male).

A sensitivity power analysis using G*Power (Faul et al., 2007) indicated that our final sample provided 80% power to detect an effect size of at least d = 0.35 for a one-tailed difference test between two independent groups with an alpha level of .05. We recognize that this effect is larger than the condition effect on behavioural intention observed in Study 3, but the sample size was limited in Study 4 by budgetary constraints.

Procedure

The study was conducted at Ludwig-Maximilians University of Munich. German residents were recruited using the laboratory's database (Greiner, 2015) and received a variable payment depending on their decisions in the game (average payment €12). The study was programmed and conducted with the software *oTree* (Chen et al., 2016).

In our version of the public goods game, the game consisted of two rounds. One player was randomly assigned as an impartial referee in each team of five, who would administer a penalty to players acting selfishly after round 1 (defined as anything less than full cooperation; Fehr & Gächter, 2000). Participants were told about the possibility of punishment, so they did not feel tricked into behaving egoistically, but they were told the punishment would be small. After the punishment, participants played a second round of the public goods game with the same team members. Participants were told that rules would remain the same in the second round except there would be no punishment. We did this to avoid a situation where participants all fully cooperated out of fear of punishment (purely extrinsic motivation). After completing round two, participants completed motive attribution and acceptance measures.

Sanctioning system manipulation

Points deducted from participants as punishment for selfish behaviour in round one were either redistributed to all players in the team (group-profit condition; reflecting relationship-oriented punishment)

or kept by the referee (self-profit condition; reflecting harm/self-oriented punishment). The size of the penalty was the same across both conditions: six points. To ensure participants attributed intention to the punishment, they were informed that the referee had multiple options and freely chose to allocate the points in this way (this was true; but unbeknownst to punishees, punishment was strongly incentivized). Deception is explicitly not permitted in this laboratory; therefore, participants would have likely accepted this information as genuine.

Materials

Outcome measures

We measured transgressors' acceptance of the punishment using a German translation of the six-item scale from Study 3 ($\alpha = .89$), and we measured prosocial behaviour by looking at participants' contributions to the public good in the second round of the game.

Punishment motive attributions

Punishment motive attributions were measured next, using the 20 items measured in Study 3, translated into German using a back-translation procedure (Brislin, 1970) involving two bilingual members of the research team. Only one item required substantive change to match the new context. Confirmatory factor analysis indicated the five-factor model using all 20 items had good fit, and scale reliabilities were acceptable: relationship-oriented motives ($\alpha = .79$); harm-oriented motives ($\alpha = .88$), self-oriented motives ($\alpha = .75$)[2]; victim-oriented motives ($\alpha = .85$); society-oriented motives ($\alpha = .71$).

Exclusion criteria

As specified in our pre-registration protocol, participants who failed two comprehension check questions—designed to test their understanding of the game structure—more than four times were excluded from analysis. In addition, we excluded participants who failed a reading check regarding the punishment manipulation.[3]

Results

Descriptive statistics and bivariate correlations for outcome measures are presented in Table 8. As expected, participants were more likely to attribute punishment to relationship-oriented motives in the group-profit condition than in the self-profit condition, $t(184.90) = 3.70$, $p < .001$, $d = 0.51$. Participants in the self-profit condition were more likely to attribute punishment to self-oriented motives, $t(197) = -4.85$, $p < .001$, $d = 0.69$, and while the mean score for this group was also higher for harm-oriented motives, the difference fell outside the bounds of statistical significance, $t(197) = -1.81$, $p = .071$, $d = 0.26$. Therefore, the self-profit system was not clearly harm-oriented, as intended, though it was more self-oriented and less relationship-oriented than the group-profit system.

Our hypothesis that relationship-oriented punishment would result in more acceptance was supported: a one-tailed *t*-test indicated that participants in the group-profit sanctioning condition accepted their punishment more than those in the self-profit sanctioning condition, $t(197) = 3.67$, $p < .001$, $d = 0.52$. Providing further support for the hypothesis, (measured) relationship-oriented attributions positively predicted acceptance ($B = 0.44$, $SE = 0.07$, $\beta = 0.41$, $p < .001$), while harm- ($B = -0.51$, $SE = 0.06$,

[2]We had anticipated that harm- and self-oriented motives to be so highly correlated in this context that they could be combined into a single scale. However, this was not the case relative to other intercorrelations (see Table 9), and results of the factor analysis indicated that a five-factor model fit the data better than a four-factor model (see Supplementary Materials). Therefore, we deviated from our pre-registration plan and maintained these as two separate scales.

[3]We forgot to describe this second check in our pre-registration plan. Only 13 participants failed this second check, and including them does not substantively change the results of the analyses.

TABLE 8 Descriptive statistics and correlations (Study 4).

	M (*SD*)		*r*					
Dependent variable	**Group-profit (*n* = 102)**	**Self-profit (*n* = 97)**	**2**	**3**	**4**	**5**	**6**	**7**
Motive attributions								
1 Relationship	2.80 (0.81)	2.33 (1.00)	−.21*	−.31**	.53**	.60**	.41**	.05
2 Harm	2.08 (0.99)	2.35 (1.11)	1	.59**	−.16*	−.30**	−.55**	−.02
3 Self	2.54 (0.95)	3.21 (0.99)		1	−.35**	−.44**	−.54**	−.01
4 Victim	3.75 (0.81)	2.53 (1.03)			1	.61**	.38**	−.14
5 Society	3.69 (0.70)	3.13 (0.90)				1	.54**	−.05
6 Acceptance	3.22 (1.02)	2.72 (0.91)					1	−.03
7 Prosocial behaviour	6.33 (5.29)	8.36 (5.80)						1

*$p < .01$.
**$p < .001$.

TABLE 9 Multivariate regression analysis (Study 4).

	Acceptance			**Prosocial behaviour**		
Predictor	**β**	***SE***	***p***	**β**	***SE***	***p***
Relationship	.10	0.07	.133	.19	0.55	.043
Harm	−.33	0.06	<.001	.002	0.47	.986
Self	−.18	0.07	.009	−.06	0.52	.528
Victim	.04	0.06	.593	−.22	0.48	.018
Society	.28	0.09	<.001	−.05	0.67	.601

β = −0.55, $p < .001$), and self-oriented attributions negatively predicted acceptance ($B = -0.53$, $SE = 0.06$, β = −0.54, $p < .001$) in bivariate regression models.

While the results for our first outcome variable (i.e. acceptance) were in line with our theorizing, the results for our second outcome variable (i.e. prosocial behaviour) were not: participants in the group-profit sanctioning condition contributed *less* to the public good in round 2 than those in the self-profit sanctioning condition, $t(197) = -2.58$, $p = .011$, $d = 0.37$. Here, neither (measured) relationship- ($B = 0.32$, $SE = 0.43$, β = .05, $p = .451$) harm- ($B = -0.12$, $SE = 0.38$, β = −.02, $p = .762$), nor self-oriented attributions ($B = -0.07$, $SE = 0.39$, β = −.01, $p = .850$) significantly predicted prosocial behaviour at the bivariate level.

Alternative interpretations of punishment motives across conditions

It is possible that despite our efforts to create sanctioning systems that clearly reflected relationship- and harm-/self-oriented punishment motives, participants may have interpreted the punishment as being rooted in alternate motives. This may have resulted in a punishment that conveyed multiple motives, not solely the ones which we were trying to manipulate. To test this, we first checked victim- and society-oriented motive attributions in the two sanctioning system conditions, in line with our pre-registration plan. Indeed, relative to those in the self-profit sanctioning condition, those in the group-profit sanctioning condition perceived punishment as more targeted at victims and society, respectively, $t(182.23) = 9.29$, $p < .001$, $d = 1.32$; $t(181.14) = 4.84$, $p < .001$, $d = 0.69$. The effect size for victim-oriented motives was particularly large, and in fact this was the predominant motive attribution in the group-profit sanctioning condition.

To examine possible effects of these alternate attributions on outcomes, acceptance and prosocial behaviour were regressed on all five attribution scales (see Table 9). Consistent with the bivariate

correlations, harm-oriented and self-oriented attributions predicted acceptance. Society-oriented attributions positively predicted acceptance. In addition, relationship-oriented motives (positively) predicted prosocial behaviour, consistent with our model, while victim-oriented motives (negatively) predicted prosocial behaviour. Since punishment in the group-profit system was most likely to be attributed to victim-oriented motives, this may explain the counter-prediction findings for condition effects on prosocial behaviour. We further explore this idea next.

Exploratory analyses

One feature of the group-profit sanctioning system condition is that the points deducted from participants as punishment were redistributed to all other players in the team. Critically, this could have been viewed as a form of compensation for the harm caused, consistent with the finding that the group-profit punishment was perceived as victim-oriented. Given the measure of prosocial behaviour was also points that were, ultimately, distributed to other players, it is possible that any reluctance by participants in the group-profit condition to contribute to the pot in round 2 may have been because some felt they had already made amends for their selfish behaviour through the reallocation of points to their team members that served as their punishment for round 1.

It is thus possible to distinguish two pathways to prosocial behaviour: a positive one via relationship-oriented attributions, and a negative one via perceptions that other team members were compensated by the punishment. One way to test this idea is by looking at mediator variables (i.e. items measuring participants' inferred punishment motives) in more detail. In fact, one item from the victim-oriented attributions scale directly taps into the perception of punishment as compensatory (in English: 'to compensate the other team members for the harm done'). Statistically controlling for participants' endorsement of this item should yield the hypothesised indirect effect of sanctioning system on prosocial behaviour via relationship-oriented motive attributions.

To test this, we ran a parallel mediation model using the PROCESS macro for SPSS (Hayes, 2013, model 4; 10,000 bootstraps). Consistent with our prediction, results revealed that participants in the group-profit condition contributed 0.45 points more to the shared pot in round 2 ($SE = 0.28$, $CI_{95\%} = [.002, 1.13]$) than those in the self-profit condition, via perceived relationship-oriented motives. A larger, opposing, indirect effect was observed via the punishment-as-compensation motive attribution item, $B = -1.07$, $SE = 0.57$, $CI_{95\%} = [-.06, -2.26]$. The compensation hypothesis, therefore, appears to be a viable explanation for why we did not observe expected effects on prosocial behaviour.

GENERAL DISCUSSION

The present research contributes to the study of transgressor perspectives of, and reactions to, punishment, and adds to a growing body of literature emphasizing the social and relational dimensions of decision-making processes. In particular, we demonstrate that motive attributions are a key dimension of transgressors' conceptualizations of punishment and are critical in fostering prosocial outcomes, particularly for influencing prosocial attitudes. We found relatively consistent results across three different methodological approaches: (i) a real-life cross-sectional survey of people interacting with criminal justice authorities; (ii) two different experimental hypothetical vignettes examining institutional infractions; and (iii) a lab-based behavioural experiment. Our findings have implications for the ways that sanctions are delivered to transgressors.

Theoretical and practical implications

Our work consolidates and expands upon various theoretical strands in the existing literature that help to answer the question of when punishment works, and when it does not. First, we provide empirical evidence that subjective understandings of punishment are key in determining outcomes. Rather than simply attending to the severity (i.e. outcome) of a punishment, transgressors respond strongly to the

communicative dimensions of punishment (Gollwitzer et al., 2011; Sarin et al., 2021), as implied, here, through punishment delivery. Specifically, we find that perceived motives for punishment are a key mechanism driving reactions to punishment. Our research indicates that transgressors respond differentially to punishment depending on the extent to which it is seen as an attempt to restore the social relationships breached by the wrongdoing (relationship-oriented motives) or motivated by spiteful or selfish reasons (harm- and self-oriented motives). This might explain why fair procedures lead to feelings of belongingness, as per procedural justice theory (van Prooijen et al., 2008). It also supports the idea that punishment can convey messages of social inclusion (a core tenet of reintegrative shaming theory; Braithwaite, 1989). Taken together, our studies suggest that motive attributions influence transgressor attitudes—while there was weaker evidence in relation to how they might alter behaviour.

In addition, we delineate two motives for punishment that are *not* directed at the transgressor: punishment that serves the victim, and punishment that serves society more broadly. Transgressors' perception that the punishment was directed towards victims was associated with negative outcomes at times (Studies 1 and 4). When punishment was viewed as a way to compensate victims for what they had lost, participants felt little need to make further contributions to the group (Study 4). It could be that when transgressors view a sanction in a 'business' frame, as a way to simply offset harm, it might not lead to affirmation of norms and behavioural change (Mulder, 2009). In addition, our findings indicate that there may be some benefit of perceiving the punishment as targeted at society (Studies 1, 2, and 4). This suggests that conveying the societal benefits of punishment can promote prosocial outcomes, though perhaps less consistently than relationship-oriented motives (Study 4).

Our findings suggest that if one's goal is to maximize prosocial outcomes, authorities should aim to convey relationship-oriented motives as much as possible when delivering sanctions. We speculate that while one of the most effective ways to influence such attributions may be through fair treatment (e.g. procedural justice), some scholars have suggested that even procedurally just treatment might, in some cases, be seen as motivated by malevolent motives (Cherney & Murphy, 2011). Therefore, it is worth considering other ways to convey desired motives. As an example, our research suggests that punishment motives can be implicit in the type or form of punishment inflicted (i.e. by redistributing the penalty for selfish behaviour to the referee vs. the group in Study 4).

Our work also has implications regarding the explicit communication of motives. When punishing others, we often tell others why we are punishing them. For example, in legal contexts, it is often customary for judges to provide justifications for a sanction during sentencing. For example, consider the following judge's remark: 'This vile conduct … calls for punishment to represent in part a retribution for these fundamental wrongs on behalf of the victim' (Warner et al., 2017). The judge is communicating a retributive goal of punishment that is victim- and potentially harm-oriented in nature. While conveying these goals might serve other desired punishment goals (e.g. validating the victim), our findings suggest that such messages are unlikely to be constructive in reforming people. Retributive messages—in courts and beyond—might rather be formulated using relationship-oriented language (e.g. imposition of suffering might pave the way to forgiveness).

In addition, future work could more purposively examine how the motives in our model map onto traditional punishment aims (e.g. retribution, deterrence). It would be interesting to explore whether our model, which was developed from the perspective of the transgressor (i.e. how punishment addresses their relational needs), compares to models developed from the perspective of those delivering or observing punishment.

Limitations

The present research is not without its limitations. Most critically, while the link between motive attributions and prosocial attitudes was robust, the relationship between attributions and prosocial behaviour was more tenuous. Though we found correlational evidence for this latter association, causal effects on prosocial behaviour were either small (Study 3) or indirect (Study 4). These results suggest that it may be difficult to reliably promote positive behaviour through punishment—the very problem that has stymied the field for decades (Sherman, 2017). Our results, though, are encouraging given the consistent positive effects of motives on

prosocial attitudes, which are a starting point for more tangible change. Moreover, in the context of repeated interactions with authorities, small effects could accumulate over time (Götz et al., 2022); a similar point has been made in explaining why single procedural justice studies rarely produce significant effects (Nagin & Telep, 2020). It may also be that effects could be larger for some populations, such as those with more hostile attributions to begin with. On the other hand, effects may be weaker in other contexts, for example, among those who prefer an autocratic style of authority (Tyler et al., 2000). Inconsistencies between our studies could also be a reflection of differences in methodologies. Further research with larger, more diverse samples and more complex designs is needed to better understand the factors that promote and inhibit model effects.

Last, it is worth considering the longevity of initial attributions. While our research indicates that transgressors' subjective construal of punishment influence attitudes immediately following sanctioning, subsequent negative experiences with authorities or experiences of the sanctions themselves may further affect their views. For example, Bullock and Bunce (2020) found that, failing to get support, prisoners viewed authorities' ostensible promotion of rehabilitation as a goal of prison as hollow and self-serving, leading to disillusionment and dashed perceptions that it was possible to achieve meaningful change. Therefore, any work that authorities do in conveying relationship-oriented motivations for punishment when delivering sanctions must ultimately be supported by actions that follow through on these sentiments.

Conclusion

This research demonstrates that the way that transgressors 'make sense' of their punishment vis-à-vis its relational implications influences attitudes and behaviour. In particular, the extent to which punishment is effective—promoting prosocial outcomes—is tied to attributions about the intention underlying an expression of punishment. Therefore, authorities and institutions may do well to direct some of their focus away from the severity of sanctions towards a careful consideration of the messages they communicate to transgressors about the purpose of their punishment.

AUTHOR CONTRIBUTIONS

Melissa de Vel-Palumbo: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Writing – original draft; Writing – review & editing. **Mathias Twardawski:** Formal analysis; funding acquisition; investigation; methodology; writing – original draft; writing – review and editing. **Mario Gollwitzer:** Conceptualization; funding acquisition; investigation; methodology; resources; writing – original draft; writing – review and editing.

ACKNOWLEDGMENTS

We kindly thank the Munich Experimental Laboratory for Economic and Social Sciences (MELESSA) for providing laboratory resources, David Greisberger for assistance with programming, and Stephan Nuding and Nadja Born for assistance with data collection.

FUNDING INFORMATION

This research was funded by a Marie Skłodowska-Curie Individual Fellowship of the European Union (EU) under the Horizon 2020 programme (839639 – PUNISH). Study 1 was supported by a LMUexcellent grant, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder.

CONFLICT OF INTEREST STATEMENT

All authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

All data and study materials, including datasets, codebooks and analysis syntax, are publicly available at https://doi.org/10.17605/OSF.IO/3TAYF.

ORCID

Melissa de Vel-Palumbo https://orcid.org/0000-0003-4765-4207
Mathias Twardawski https://orcid.org/0000-0003-0543-277X
Mario Gollwitzer https://orcid.org/0000-0003-4310-4793

REFERENCES

Ahmed, E., & Braithwaite, V. (2004). "What, me ashamed?" shame management and school bullying. *Journal of Research in Crime and Delinquency*, *41*(3), 269–294. https://doi.org/10.1177/0022427804266547

Balafoutas, L., García-Gallego, A., Georgantzis, N., Jaber-Lopez, T., & Mitrokostas, E. (2020). Rehabilitation and social behavior: Experiments in prison. *Games and Economic Behavior*, *119*, 148–171. https://doi.org/10.1016/j.geb.2019.10.009

Beijersbergen, K. A., Dirkzwager, A. J. E., Eichelsheim, V. I., Van der Laan, P. H., & Nieuwbeerta, P. (2015). Procedural justice, anger, and prisoners' misconduct: A longitudinal study. *Criminal Justice and Behavior*, *42*(2), 196–218. https://doi.org/10.1177/0093854814550710

Bies, R. J. (2001). Interactional (in) justice: The sacred and the profane. In J. Greenberg & R. Cropanzano (Eds.), *Advances in organizational justice* (pp. 89–118). Stanford University Press.

Bies, R. J., & Moag, J. S. (1986). Interactional justice: Communication criteria of fairness. In R. J. Lewicki, B. H. Sheppard, & M. H. Bazerman (Eds.), *Research on negotiation in organizations* (pp. 43–55). JAI Press.

Blader, S. L., & Tyler, T. R. (2003). A four-component model of procedural justice: Defining the meaning of a "fair" process. *Personality and Social Psychology Bulletin*, *29*(6), 747–758. https://doi.org/10.1177/0146167203252811

Blader, S. L., & Tyler, T. R. (2015). Relational models of procedural justice. In R. S. Cropanzano & M. L. Ambrose (Eds.), *The Oxford handbook of justice in the workplace* (2nd ed., pp. 351–369). Oxford University Press.

Bobocel, D. R., & Gosse, L. (2015). Procedural justice: A historical review and critical analysis. In R. S. Cropanzano & M. L. Ambrose (Eds.), *The Oxford handbook of justice in the workplace* (pp. 51–87). Oxford University Press.

Bobocel, D. R., & Holmvall, C. M. (2001). Are interactional justice and procedural justice different. In S. Gilliland, D. Steiner, & D. Skarlicki (Eds.), *Theoretical and cultural perspectives on organizational justice* (pp. 85–108). Information Age.

Bottoms, A., & Tankebe, J. (2012). Beyond procedural justice: A dialogic approach to legitimacy in criminal justice. *Journal of Criminal Law and Criminology*, *102*, 119–170.

Braithwaite, J. (1989). *Crime, shame and reintegration*. Cambridge University Press.

Brandt, H., Hauert, C., & Sigmund, K. (2006). Punishing and abstaining for public goods. *Proceedings of the National Academy of Sciences*, *103*(2), 495–497. https://doi.org/10.1073/pnas.0507229103

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, *1*(3), 185–216. https://doi.org/10.1177/135910457000100301

Brockner, J., & Wiesenfeld, B. M. (1996). An integrative framework for explaining reactions to decisions: Interactive effects of outcomes and procedures. *Psychological Bulletin*, *120*(2), 189–208. https://doi.org/10.1037/0033-2909.120.2.189

Bullock, K., & Bunce, A. (2020). 'The prison don't talk to you about getting out of prison': On why prisons in England and Wales fail to rehabilitate prisoners. *Criminology & Criminal Justice*, *20*(1), 111–127. https://doi.org/10.1177/1748895818800743

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

Chebat, J.-C., & Slusarczyk, W. (2005). How emotions mediate the effects of perceived justice on loyalty in service recovery situations: An empirical study. *Journal of Business Research*, *58*(5), 664–673. https://doi.org/10.1016/j.jbusres.2003.09.005

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97. https://doi.org/10.2139/ssrn.2806713

Cherney, A., & Murphy, K. (2011). Understanding the contingency of procedural justice outcomes. *Policing*, *5*(3), 228–235. https://doi.org/10.1093/police/par030

Colquitt, J. A. (2001). On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology*, *86*(3), 386–400. https://doi.org/10.1037/0021-9010.86.3.386

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

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.4324/9780203127698

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4), 980–994. https://doi.org/10.1257/aer.90.4.980

Frijda, N. H. (1994). The lex talionis: On vengeance. In S. H. M. van Goozen, N. E. van der Poll, & J. A. Sergeant (Eds.), *Emotions: Essays on emotion theory* (pp. 263–289). Erlbaum.

Gollwitzer, M., Meder, M., & Schmitt, M. (2011). What gives victims satisfaction when they seek revenge? *European Journal of Social Psychology*, *41*, 364–374.

Gollwitzer, M., & Okimoto, T. G. (2021). Downstream consequences of post-transgression responses: A motive-attribution framework. *Personality and Social Psychology Review*, *25*, 275–294.

Götz, F. M., Gosling, S. D., & Rentfrow, P. J. (2022). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives on Psychological Science*, *17*(1), 205–215. https://doi.org/10.1177/1745691620984483

Greenberg, J. (1993). Stealing in the name of justice: Informational and interpersonal moderators of theft reactions to underpayment inequity. *Organizational Behavior and Human Decision Processes*, *54*(1), 81–103. https://doi.org/10.1006/obhd.1993.1004

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114–125. https://doi.org/10.1007/s40881-015-0004-4

Griffin, M. L. (2006). Penal harm and unusual conditions of confinement: Inmate perceptions of 'hard time' in jail. *American Journal of Criminal Justice*, *30*(2), 209–226. https://doi.org/10.1007/bf02885892

Grootelaar, H. A. M., & van den Bos, K. (2018). How litigants in Dutch courtrooms come to trust judges: The role of perceived procedural justice, outcome favorability, and other sociolegal moderators. *Law & Society Review*, *52*(1), 234–268. https://doi.org/10.1111/lasr.12315

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford.

Komorita, S. S., & Parks, C. D. (1994). *Social dilemmas*. Westview Press.

Kurz, T., Thomas, W. E., & Fonseca, M. A. (2014). A fine is a more effective financial deterrent when framed retributively and extracted publicly. *Journal of Experimental Social Psychology*, *54*, 170–177. https://doi.org/10.1016/j.jesp.2014.04.015

Leventhal, G. S. (1980). What should be done with equity theory? In K. J. Gergen, M. S. Greenberg, & R. H. Willis (Eds.), *Social exchange: Advances in theory and research* (pp. 27–55). Springer. https://doi.org/10.1007/978-1-4613-3087-5_2

Lind, E. A., & Tyler, T. R. (1988). *The social psychology of procedural justice*. Springer.

Lotz, S., Okimoto, T. G., Schlösser, T., & Fetchenhauer, D. (2011). Punitive versus compensatory reactions to injustice: Emotional antecedents to third-party interventions. *Journal of Experimental Social Psychology*, *47*(2), 477–480. https://doi.org/10.1016/j.jesp.2010.10.004

McGrath, A. (2009). Offenders' perceptions of the sentencing process: A study of deterrence and stigmatisation in the New South Wales Children's court. *Australian & New Zealand Journal of Criminology*, *42*(1), 24–46. https://doi.org/10.1375/acri.42.1.24

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/e518362013-127

Mulder, L. B. (2009). The two-fold influence of sanctions on moral norms. In *Psychological perspectives on ethical behavior and decision making* (pp. 169–180). Information Age Publishing.

Nagin, D. S., & Telep, C. W. (2020). Procedural justice and legal compliance: A revisionist perspective. *Criminology & Public Policy*, *19*(3), 761–786. https://doi.org/10.1111/1745-9133.12499

Okimoto, T. G., & Wenzel, M. (2009). Punishment as restoration of group and offender values following a transgression: Value consensus through symbolic labelling and offender reform. *European Journal of Social Psychology*, *39*(3), 346–367. https://doi.org/10.1002/ejsp.537

Okimoto, T. G., & Wenzel, M. (2011). Third-party punishment and symbolic intragroup status. *Journal of Experimental Social Psychology*, *47*(4), 709–718. https://doi.org/10.1016/j.jesp.2011.02.001

Orth, U. (2003). Punishment goals of crime victims. *Law and Human Behavior*, *27*(2), 173–186. https://doi.org/10.1023/a:1022547213760

Oswald, M. E., Hupfeld, J., Klug, S. C., & Gabriel, U. (2002). Lay-perspectives on criminal deviance, goals of punishment, and punitivity. *Social Justice Research*, *15*(2), 85–98. https://doi.org/10.1023/A:1019928721720

Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to power analysis for simple experimental designs. *International Review of Social Psychology*, *31*(1), 20. https://doi.org/10.5334/irsp.181

Prochaska, J. O., Redding, C. A., & Evers, K. E. (2015). The transtheoretical model and stages of change. In K. Glanz, B. K. Rimer, & K. Viswanath (Eds.), *Health behavior: Theory, research, and practice* (5th ed., pp. 125–148). Jossey-Bass.

Reisig, M. D., Tankebe, J., & Mesko, G. (2014). Compliance with the law in Slovenia: The role of procedural justice and police legitimacy. *European Journal on Criminal Policy and Research*, *20*(2), 259–276. https://doi.org/10.1007/s10610-013-9211-9

Sarin, A., Ho, M. K., Martin, J. W., & Cushman, F. A. (2021). Punishment is organized around principles of communicative inference. *Cognition*, *208*, 104544.

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science*, *8*(4), 379–386. https://doi.org/10.1177/1948550617715068

Sherman, L. W. (1993). Defiance, deterrence, and irrelevance: A theory of the criminal sanction. *Journal of Research in Crime and Delinquency*, *30*(4), 445–473. https://doi.org/10.1177/0022427893030004006

Sherman, L. W. (2017). Experiments in criminal sanctions: Labeling, defiance, and restorative justice. In *Labeling theory* (pp. 149–176). Routledge.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, *89*(6), 845–851. https://doi.org/10.1037/0022-3514.89.6.845

Tomlinson, E. C. (2012). The impact of apologies and promises on post-violation trust: The mediating role of interactional justice. *International Journal of Conflict Management*, *23*(3), 224–247. https://doi.org/10.1108/10444061211248930

Tyler, T. R. (2003). Procedural justice, legitimacy, and the effective rule of law. *Crime and Justice*, *30*, 283–357. https://doi.org/10.1086/652233

Tyler, T. R. (2008). Procedural justice and the courts. *Court Review: The Journal of the American Judges Association*, *44*(1/2), 26–31.

Tyler, T. R., & Bies, R. J. (2015). Beyond formal procedures: The interpersonal context of procedural justice. In J. S. Carroll (Ed.), *Applied social psychology and organizational settings* (pp. 77–98). Psychology Press.

Tyler, T. R., & Blader, S. L. (2000). *Cooperation in groups: Procedural justice, social identity, and behavioral engagement*. Psychology Press.

Tyler, T. R., Lind, E. A., & Huo, Y. J. (2000). Cultural values and authority relations: The psychology of conflict resolution across cultures. *Psychology, Public Policy, and Law*, *6*(4), 1138–1163. https://doi.org/10.1037//1076-8971.6.4.1138

Tyler, T. R., & Trinkner, R. (2017). *Why children follow rules: Legal socialization and the development of legitimacy*. Oxford University Press.

Tyler, T. R., & Wakslak, C. J. (2004). Profiling and police legitimacy: Procedural justice, attributions of motive, and acceptance of police authority. *Criminology*, *42*(2), 253–282. https://doi.org/10.1111/j.1745-9125.2004.tb00520.x

Van Lange, P. A. M. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology*, *77*(2), 337–349. https://doi.org/10.1037/0022-3514.77.2.337

van Prooijen, J.-W. (2017). *The moral punishment instinct*. Oxford University Press.

van Prooijen, J.-W., Gallucci, M., & Toeset, G. (2008). Procedural justice in punishment systems: Inconsistent punishment procedures have detrimental effects on cooperation. *British Journal of Social Psychology*, *47*(2), 311–324.

Vidmar, N., & Miller, D. T. (1980). Socialpsychological processes underlying attitudes toward legal punishment. *Law and Society Review*, *14*(3), 565–602. https://doi.org/10.2307/3053193

Warner, K., Davis, J., & Cockburn, H. (2017). The purposes of punishment: How do judges apply a legislative statement of sentencing purposes? *Criminal Law Journal*, *41*(2), 69–85.

Woodyatt, L., & Wenzel, M. (2013). The psychological immune response in the face of transgressions: Pseudo self-forgiveness and threat to belonging. *Journal of Experimental Social Psychology*, *49*(6), 951–958. https://doi.org/10.1016/j.jesp.2013.05.016

Woodyatt, L., Wenzel, M., & de Vel-Palumbo, M. (2017). Working through psychological needs following transgressions to arrive at self-forgiveness. In L. Woodyatt, E. L. Worthington, Jr., M. Wenzel, & B. J. Griffin (Eds.), *Handbook of the psychology of self-forgiveness* (pp. 43–58). Springer.

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493–504. https://doi.org/10.1037/pspa0000056

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: de Vel-Palumbo, M., Twardawski, M., & Gollwitzer, M. (2023). Making sense of punishment: Transgressors' interpretation of punishment motives determines the effects of sanctions. *British Journal of Social Psychology*, *62*, 1395–1417. https://doi.org/10.1111/bjso.12638