Undoing harm:
The communicative content of action-oriented and person-oriented punishment

Christian Mott[1] & Larisa Heiphetz Solomon[2]

[1] Department of Psychology, Columbia University. 1190 Amsterdam Ave., New York, NY 10027, United States. E-mail: christian.mott@columbia.edu. Phone: 212-853-1406. (Corresponding author)

[2] Department of Psychology, Columbia University. 1190 Amsterdam Ave., New York, NY 10027, United States. E-mail: lah2201@columbia.edu. Phone: 212-854-1348.

Author Note: This project was made possible through the support of NSF CAREER grant #2044360 and funding from Columbia University, both awarded to the second author. Any opinions expressed are those of the authors alone and should not be construed as representing the opinions of any organizations that provided support for this project. The authors wish to thank the members of the Columbia Social and Moral Cognition Lab for providing feedback. Vignettes, measures, materials, data, and analysis code for all studies appear at: https://osf.io/nxw24/?view_only=b92f21261c8e4804857bccd6099af96b.

Abstract

Punishment can serve as a form of communication: People use punishment to express information to its recipients and interpret punishment between third parties as having communicative content. Prior work on the expressive function of punishment has primarily investigated the capacity of punishment in general to communicate a single type of message – e.g., that the punished behavior violated an important norm. The present work expands this framework by testing whether different types of punishment communicate different messages. We distinguish between *person-oriented* punishments, which seek to harm the recipient, and *action-oriented* punishments, which seek to undo a harmful action. We show that people interpret action-oriented punishments, compared to person-oriented punishments, to indicate that the recipient will change for the better (Study 1). The communicative theory can explain this finding if people understand action-oriented punishment to send a message that is more effective than person-oriented punishment at causing such a change. Supporting this explanation, inferences about future behavior track the recipients' beliefs about the punishment they received, rather than the punisher's intentions or the actual punishment imposed (Study 2). Indeed, when actual recipients of a person-oriented punishment believed they received an action-oriented punishment and vice versa, predictions of future behavior tracked the recipients' beliefs rather than reality, and judgments about what the recipients learned from the punishments mediate this effect (Study 3). Together, these studies demonstrate that laypeople think different types of punishment send different messages to recipients and that these messages are differentially effective at bringing about behavioral changes.

Keywords: punishment; moral psychology; experimental jurisprudence; social cognition

Undoing harm:

The communicative content of action-oriented and person-oriented punishment

Imagine two children, each playing with a toy. After a few moments of peaceful coexistence, one smashes the other's toy. A caregiver witnessing this scene will likely punish the misbehaving child. There are two obvious punishments available: (1) taking away the misbehaving child's toy and putting it somewhere no one can play with it, or (2) taking away the misbehaving child's toy and giving it to the other child. The first type of punishment is *person-oriented*: it targets the individual who committed the offense and makes life unpleasant for her. The second type of punishment is *action-oriented*: it targets the wrongful behavior by restoring the world as much as possible to the way it was before the action took place. In this example, the action-oriented punishment inflicts a negative consequence on the transgressor in a way that directly benefits the victim – i.e., the transgressor gives up her toy and the victim receives it. In contrast, the person-oriented punishment inflicts the same negative consequence on the transgressor but does not directly benefit the victim or bring the world closer to its pre-offense state.

These two types of punishment occur in a variety of situations. In family contexts, a parent may respond to transgressions by encouraging their child to fix their wrongdoing or by inflicting negative consequences that do not directly ameliorate the harm the child caused, as described above. In social settings, a group of friends may ostracize one member who harmed another, or they may require that member to make specific amends to the person he harmed in order to remain in the group. Within the legal system, the state can impose punishments that target a transgressor's behavior, such as requiring a thief to spend time volunteering for the organization from which she stole money, or punishments that target transgressors themselves,

such as requiring the thief to spend the same amount of time in jail. Although these responses vary in many ways, they are all types of *punishment*: In each case, an authority imposes some sort of loss – e.g., of well-being, property, or autonomy – as a response to norm violation (Brooks, 2012; Hart, 1968). The difference between the two types of punishments is whether they benefit victims and whether the punishment restores the world to its pre-offense state, in whole or in part.

When deciding how to punish third parties, people distinguish between these two types of responses. For example, people prefer to impose action-oriented over person-oriented punishment (Bicchieri & Maras, 2022; Riedl et al., 2015), except when they think the norm violation results from the perpetrator being a fundamentally bad person (Maffly-Kipp et al., 2022). People also prefer action-oriented punishment when they want to ensure that recipients recognize that the harm is, in fact, a punishment for a particular behavior – rather than an unconnected harm by the punisher – and thereby potentially learn not to perform that behavior in the future (Sarin et al., 2021). This research suggests that people understand punishment as communicative, that they perceive action-oriented and person-oriented punishments to communicate different messages to the punishment recipient, and that they believe action-oriented punishment sends a message that is more effective at teaching the punishment recipient to behave normatively in the future.

To test these possibilities, we investigated the inferences people draw when they learn that someone has received action-oriented or person-oriented punishment from the state after a criminal conviction. In this context, where there is no doubt that the harms the state imposes are punishments, we explored whether laypeople nevertheless perceive one type of punishment as more effective than the other at teaching recipients how to behave. Three studies found a

significant difference in expectations about future behavior between individuals who had committed the same crime but received different types of punishment. These studies also suggest that this difference arises from a perception that action-oriented and person-oriented punishments send different messages.

Theories of Punishment

The distinction between action-oriented and person-oriented punishment is largely independent of the traditional distinctions between different theories[1] of punishment, such as retributivism, specific and general deterrence, rehabilitation, and expressivism (Brooks, 2012; Hart, 1968; Mott & Solomon, 2024). These traditional theories of punishment differ in the *ultimate goals* they propose punishment does or should serve. For example, a retributivist imposes punishment to give the recipient what they deserve, conceptualized as a harm proportionate to the wrongdoing, the infliction of which the retributivist views as an end in itself (Husak, 2017). The distinction between action-oriented and person-oriented punishment, by contrast, is a difference in *instrumental goals*. A person inflicting a person-oriented punishment seeks to achieve the instrumental goal of harming the wrongdoer, which may, in turn, contribute to an ultimate goal of, e.g., retribution or deterrence. A person inflicting an action-oriented punishment harms a wrongdoer to achieve the instrumental goal of addressing the harmful effects of the wrongdoing, which may, in turn, contribute to an ultimate goal of, e.g., the wrongdoer's rehabilitation

[1] In philosophical work, these theories of punishment arise as potential normative justifications for punishment (Brooks, 2012). In psychological work, these theories are goals punishers may seek to achieve through punishment (Carlsmith, 2008; Carlsmith et al., 2002; Darley et al., 2000). We refer to the latter forms of these theories.

Thus, a punisher could use either type of punishment to accomplish many possible ultimate goals. Either punishment can serve the ultimate goal of retribution if the harm is proportional to the norm violation (Husak, 2017) or the goal of deterrence if the harm raises the perceived cost of the norm violation enough to outweigh its perceived benefit (Brooks, 2012). Likewise, both types of punishment could conceivably help rehabilitate recipients; whether they actually do so is an empirical question. Both action-oriented and person-oriented punishments can also serve the goals of a hybrid theory of punishment like the communicative theory – that is, the theory that a core function of punishment is to send a message to the punishment recipient and the community about the recipient's violation of the community's values (Brooks, 2012; Duff, 2001; Dunlea & Heiphetz, 2021; Sarin et al., 2021). Indeed, we argue that these two types of punishment both express messages – they just express different messages.

Thus, although "person-oriented punishment" may sound like another name for retributive justice, the two categories apply to different types of goals. Whether or not a punishment qualifies as a person-oriented punishment depends on its instrumental goal – if it harms the recipient but does not address the harm to the victim, it is person-oriented; if it harms the recipient and ameliorates the effects of their wrong, it is action-oriented. By contrast, whether a punishment qualifies as retributive depends on its ultimate goal: According to retributive theories of punishment, imposing deserved harm on a wrongdoer is an end in itself (Caruso, 2021), and retributive justice therefore encompasses punishment imposed for the ultimate goal of achieving that end (Brooks, 2012). Person-oriented punishment and retributive justice are not entirely unrelated, however. Retributivists will often use person-oriented punishment because their ultimate goal does not require undoing the harmful effect of the wrongful conduct. It is, however, possible to imagine a scenario in which an action-oriented punishment would inflict

greater harm on a punishment recipient than a person-oriented punishment of similar magnitude, in which case a retributivist may prefer an action-oriented punishment. As an example, imagine a person whose motivation to steal from a particular victim stemmed both from a desire for the stolen property and from a animus towards the particular victim. That person would likely suffer greater harm from being forced to provide restitution to the victim than from paying a fine of equal size to the state.

Likewise, although action-oriented punishments bear a conceptual resemblance to restorative justice in that they both attempt to mitigate the negative consequences of transgressions for victims, there are also important differences between these two responses to norm violations. Like action-oriented punishment, restorative justice approaches emphasize the importance of outcomes that benefit the victim and, if possible, restore the relationships between transgressors, their victims, and the broader community (Braithwaite, 2002; Brooks, 2012; Zehr, 1990). However, legal theorists generally conceptualize restorative justice as an alternative to punishment when responding to criminal offenses (Boonin, 2008; Brooks, 2012). Advocates of restorative justice emphasize its capacity to heal wrongdoing and repair broken relationships *without* inflicting punishment (Gibson, 2021; Sered, 2019). For this reason, restorative justice programs generally exist as alternatives to the criminal justice system, particularly for minor offenses. Instead of proceeding before a judge who imposes a sentence, common restorative justice programs center on a conference between transgressors and their victim(s) in order to decide upon a series of actions for the transgressor to take, which can be memorialized in a contract (Brooks, 2012).

By contrast, action-oriented punishments are *punishments* – i.e., responses to norm violations that one person coercively imposes on another – rather than the result of any type of

agreement. This coercive feature of action-oriented punishments brings them within the standard definition of punishment. And when a person in a position of authority imposes an action-oriented punishment on a recipient, this type of punishment, like other types of punishment, can also potentially possesses communicative content.

Likewise, the distinction between action-oriented and person-oriented punishment is distinct from the taxonomy of punishment motivations introduced by de Vel-Palumbo and colleagues (2023a). The five motivations in that taxonomy – relationship-oriented motives, harm-oriented motives, self-oriented motives, victim-oriented motives, and society-oriented motives – correspond to different ultimate goals of punishment. A punisher could use either action-oriented or person-oriented punishment to accomplish several of these goals, depending on the context. Indeed, in one study (de Vel Palumbo et al., 2023a, Study 4), in which some participants received a punishment we would describe as action-oriented (transferring hoarded tokens to other players) and other participants received a punishment we would describe as person-oriented (transferring hoarded tokens to the punisher), participants attributed all five motivations to some degree in both conditions.

These findings support the idea that the distinction between action-oriented and person-oriented punishment picks out a different way punishments can vary than these distinctions between ultimate goals of punishment, though the two may be correlated. Recipients of the action-oriented punishment were significantly more likely to attribute relationship-oriented, victim-oriented, and society-oriented motivations to the punisher, while recipients of the person-oriented punishment were significantly more likely to attribute self-oriented motivations and marginally more likely to attribute harm-oriented motivations (de Vel-Palumbo et al., 2023a). As

we will discuss below, these findings are also consistent with the claim that action-oriented and person-oriented punishments can communicate different messages to recipients.

Communicating with Action-Oriented and Person-Oriented Punishment

The communicative theory of punishment originated in legal philosophy as a normative justification for the practice of punishment and as an alternative to justifications like retribution, deterrence, and incapacitation (Brooks, 2012; Nahmias & Ahroni, 2017). Although in this paper we will focus on a descriptive version of this theory (i.e., the theory that people understand punishment, at least in part, to serve communicative ultimate goals), we briefly outline features of the normative theory that, if present in the descriptive theory, provide support for the proposal that different types of punishment communicate different messages.

On the normative theory, the state is justified in harming a person in response to that person's legal violation because the harm communicates a valuable message to the recipient of the punishment and to the public. Different theorists have proposed different accounts of the content of this message (Mott & Solomon, 2024). These different contents fall into two general categories. One category of messages that theorists have argued punishment can send – punishment as an educational communication about the wrongfulness of an action (Ewing, 1943; Gahringer, 1960; Hampton, 1984) and punishment as condemnation of an action (Boonin, 2008; Duff, 2001; Primoratz, 1989) – treat the recipient of the punishment as a member of the punisher's community who must internalize the community's norms (Nahmias & Ahroni, 2017), with the expectation that the recipient will continue to be a member of that community. This message is therefore *didactic*. By contrast, another type of message that theorists have argued punishment can send – punishment as a communication of hatred towards the recipient (Feinberg, 1965; Stephen, 2014/1883) – lacks that quality. This message is *ostracizing*, in that it

distances the recipient from the community and does not provide any clear path to reintegration. On one view, punishment can send either a didactic or an ostracizing message depending on the severity of the crime and the nature of the punishment (Duff, 2001). That is, punishment does not necessarily send one type of message or the other. Instead, features of the punishment can affect what type of message it sends. This argument therefore supports the hypothesis that action-oriented and person-oriented punishments may send different messages, if the descriptive version of the communicative theory of punishment is true.

Existing research indicates that the communicative theory does provide a descriptive account of how people think about punishment, at least in some respects (Dunlea & Heiphetz, 2021; Nahmias & Ahroni, 2017). First, the communicative theory is consistent with the way people *impose* punishment. For example, punishers experience greater satisfaction when a punishment recipient acknowledges that they have been punished for a particular reason – e.g., violating a moral norm – and shows some subsequent change in moral attitude than when the recipients make no acknowledgment (Funk et al., 2014). Similarly, people prefer to impose punishment that communicates the nature of the norm violation over punishment that does not (Sarin et al., 2021) and use punishment in a way more consistent with an attempt to communicate the punisher's preferences than to reinforce a desirable behavior (Cushman et al., 2019; Ho et al., 2019).

Second, the communicative theory is consistent with the way people respond to punishments imposed on themselves. When people receive punishment, they draw inferences about the motivations of the punisher based both on the way the punishment is imposed and the type of punishment (de Vel-Palumbo et al., 2023a). These inferences, in turn, affect the degree to which people accept punishment, their motivation to change, and their actual future behavior.

Third, the communicative theory is consistent with the way people *interpret* punishment between others. In a series of studies, participants in the role of third-party observers viewed social punishment as a didactic communication that could help the recipient "get the message" that their behavior was wrong (Sarin et al., 2021, pp. 4, 9-10). Participants in these studies treated punishment like language – a form of communication in which interpretation requires using the communicator's behaviors to infer their intentions – and therefore expected a harm more consistent with an intention to punish (rather than an intention to benefit the punisher) to more effectively teach the recipient how to behave normatively in the future. Developmental research also suggests that children can understand punishment communicatively (Bregant et al., 2016).

Although this prior research has provided evidence that people can understand punishment to communicate a message with didactic content, it did not isolate individual factors that determine the message a punishment sends. Existing evidence suggests that the distinction between action-oriented and person-oriented punishment may serve as one such factor – and, specifically, that laypeople perceive action-oriented punishments to send a more didactic message than person-oriented punishments. In this existing work, researchers have not used the terms "action-oriented" and "person-oriented," but they have investigated the way people impose and understand specific punishments that fall within these categories, and their findings are consistent with this proposal.

First, this prior research suggests that both adults and children distinguish between action-oriented and person-oriented punishments, because they generally prefer to impose action-oriented rather than person-oriented punishments. Children between three and five years old commonly restore stolen items to their original owner rather than merely taking them away

from the person who committed the theft (Riedl et al., 2015).[2] Likewise, in research using an economic game paradigm, adults both compensated victims by giving them resources and punished perpetrators by taking away resources (Bicchieri & Maras, 2022). That is, they produced the same end-state as an action-oriented punishment in response to serious intentional norm violations.

Second, existing research suggests that adults think person-oriented punishments are particularly appropriate for wrongful acts that reflect an underlying bad character, suggesting they perceive person-oriented punishment to send a more ostracizing, and less didactic, message than action-oriented punishment. Although adults generally prefer to impose action-oriented rather than person-oriented punishments, this preference flips in one circumstance: When they think the transgressor's norm violation reflects their *true self* (Maffly-Kipp et al., 2022), i.e., the essence of who that person is (Christy et al., 2019; Strohminger et al., 2017; Tobia, 2016). Generally, people assume that the true selves of others are morally good, and that morally bad behaviors are merely superficial or accidental properties that do not reflect a person's essence (Lee et al., in press; Newman et al., 2015). Thus, people believe that others who improve over time come to reflect their true selves more than those who deteriorate (Heiphetz et al., 2018; Tobia, 2016). Nevertheless, people think it is possible for someone to have a morally bad true

[2] Another strand of research comparing children's judgments about person-oriented punishment with pure victim compensation – i.e., payment of money to the victim that does not come from the punishment recipient – has sometimes found a preference for punishment (McAuliffe & Dunham, 2021) and other times found a preference for compensation (Lee & Warneken, 2020). On this comparison, adults show a preference for punishment when in the role of punisher but a preference for compensation in the role of victim (FeldmanHall et al., 2014). Because these studies do not include action-oriented punishment as an option, they do not bear directly on the question under consideration here. However, there are possible communicative explanations for these preferences, including a desire by the punisher to bolster their moral reputation by communicating how seriously they take the transgression and how much they care about the victims (Heffner & FeldmanHall, 2019).

self (Newman et al., 2015), and the preference for person-oriented punishment in these circumstances (Maffly-Kipp et al., 2022) is consistent with the claim that people reserve person-oriented punishments for particularly *bad people* and action-oriented punishments for people who have merely performed *bad actions*.

The communicative theory of punishment provides a potential explanation of this finding. On this theory, one of the goals people can pursue when imposing punishment is to communicate to people who have engaged in wrongdoing that their behavior was unacceptable and that they should behave differently in the future (Sarin et al., 2021). For most recipients, whom people assume have good true selves (Strohminger et al., 2017), such communication could change their future behavior by helping them better understand how to behave in ways consistent with those true selves. When punishing these recipients, people prefer action-oriented punishments. By contrast, if individuals perceive a punishment recipient's true self as morally bad, they may conclude that they and the recipient have fundamentally different values, such that attempting to teach moral behavior to such a recipient would be futile. For these recipients, therefore, people may prefer to convey an ostracizing message, which could be the message that person-oriented punishment sends.

If people acting as punishers believe that person-oriented and action-oriented punishments have different communicative contents, they may also draw different inferences about the effects of these punishments when imposed by third parties. Specifically, because action-oriented punishments target behaviors, people may interpret them to communicate a didactic message that the punishment-eliciting *behaviors* are bad. In contrast, person-oriented punishments are less focused on the behavior and may therefore communicate either an unclear message about the reason for punishment or an ostracizing message that the recipients

themselves are bad. Thus, action-oriented punishments may be more likely than person-oriented punishments to lead an observer to infer that the recipients of the punishment will behave better in the future – either because people believe that a didactic message changes behavior more effectively than a confusing or ostracizing message or because they trust the punisher's judgment that the recipient of a person-oriented punishment is a bad person.

However, there are also reasons to predict that people may *not* expect better future behavior from recipients of action-oriented, as compared to person-oriented, punishment. First, people often infer that observable behavior reflects fixed underlying character traits (Klein & O'Brien, 2016; Uhlmann et al., 2015). In particular, people often expect that individuals who previously broke the law will reoffend (Denver et al., 2017). This tendency to infer that people who have engaged in wrongdoing will continue to do so may overwhelm any difference in communicative content between action-oriented and person-oriented punishment. Second, there is evidence that in certain contexts – specifically, where normative future behavior would benefit the very victims recently compensated by an action-oriented punishment – people who have received punishment for wrongdoing will, in fact, engage in better future behavior after receiving person-oriented punishment rather than action-oriented punishment (de Vel-Palumbo et al, 2023a). If people are aware of this tendency, and they think it generalizes outside the narrow contexts described above, then they may expect better future behavior from recipients of person-oriented punishment.

The present research provides evidence about which of these two predictions is correct. In a series of experiments, we investigated whether people draw difference inferences about a wrongdoer's future behavior based on the type of punishment received and, if so, which type of punishment they predict will lead to better future behavior.

Overview of Current Research

The three studies reported in this paper investigated punishment in the context of the legal system, where it was unambiguous that a punishment was imposed and why. Specifically, we examined whether adults expect better future behavior from people who receive action-oriented, versus person-oriented, punishments in criminal proceedings and whether this expectation arises from beliefs about the communicative function of punishment. Taken together, these data suggest that people interpret action-oriented punishments as communicating a didactic message that is more effective at producing normative behavior than the message communicated by person-oriented punishments. As a result, action-oriented punishments give rise to more optimistic inferences about future behaviors than do person-oriented punishments.

Study 1

Study 1 investigated whether participants informed about two people who committed identical crimes would predict better future behavior from a person who received an action-oriented punishment than a person who received a person-oriented punishment of the same magnitude. To reduce variation due to geographical differences in punishment practices (e.g., Hyatt et al., 2022), we recruited participants from the United States exclusively. Because extreme racial disparity is a prominent feature of the United States criminal justice system (Alexander, 2010), we also investigated whether these predictions depended on the race of the punishment recipients by pairing half the stimuli with pictures of two Black men and half the stimuli with pictures of two White men.

Methods

Participants

Based on a pilot study (see Supplementary Materials), we expected a small effect size (d = 0.2). Seeking power of 0.8 and planning to use a one-sample t-test with a standard alpha level as our primary analysis, we used G*Power 3.1 to determine that we needed a sample of at least 199 participants. To account for possible exclusions, we recruited 245 adult participants on Amazon's Mechanical Turk.[3] In accordance with the pre-registration (https://aspredicted.org/3J2_XKW), we excluded thirty-four participants who completed the entire survey in less than one minute, did not complete the survey, or failed an attention check, leaving an analysis set of 211 participants[4] (M_{age} = 39.74 years; SD_{age} = 12.12 years; 43% female, 56% male, 1% non-binary, 1 participant who did not report a gender). Based on a sensitivity analysis, this sample size allowed us to detect an effect of size d = 0.19 or greater, given a desired power of 0.8 and a standard alpha level of .05.

Participants self-identified their race in the following percentages: 83% White or European-American, 9% Asian or Asian-American, 4% Black or African-American, 4% multiracial, 0.5% Native American or Pacific Islander, and 0.5% "option not listed." We asked about ethnicity separately from race, and 11% of participants self-identified as Hispanic or Latina/o. The sample's highest level of education was distributed as follows: 12% high school or GED, 18% some college or university, 9% associate's degree, 45% bachelor's degree, 15% master's or professional degree, and 3% PhD.[5]

Procedure

[3] The following applies to all studies reported in this paper: We obtained informed consent from all participants at the beginning of the online surveys. The Columbia University Institutional Review Board reviewed and approved these studies.

[4] Here and in all subsequent studies, the patterns of results reported in the main text also emerged when analyzing data from all respondents.

[5] Demographic percentages reported throughout the paper may not add up to 100% due to rounding.

After providing informed consent, participants learned that they would read about pairs of men – referred to below as the "targets" – who had each committed the same type of crime. These men did not know each other, were not working together, and were not in the same place. The materials informed participants that across the thousands of cases brought in courts across the country on any given day, it is very common for two people in different places to have committed similar offenses.

Participants then received six vignettes in counterbalanced order. Each vignette described two men who had separately committed the same one of six offenses: identity theft, fraud, larceny, robbery, armed robbery, and burglary. These offenses all involve the wrongful taking of another's property and are common categories of property theft used in every jurisdiction across the United States. The stimuli described each offense in a way that would satisfy the statutory definition of these crimes in most jurisdictions. For example, we described a robbery in which the targets snatched the victim's bag, pulled on the bag until the strap broke, and then ran away. To ensure that the crimes did not differ substantially in severity, we described the offenses to avoid any inference that the crime created a meaningful risk of physical harm. For example, in the armed robbery vignette, the targets did not have a weapon and instead made the hands in their pockets look like weapons. That behavior can satisfy the elements of armed robbery (e.g., *State v. Chapland*, 2006) but reduces the actual level of risk to the victim. The targets in all the vignettes reported in this paper were male because criminal defendants in the United States are overwhelmingly male (Kaeble & Beatty, 2016).

Participants learned that one of the targets had received a person-oriented punishment – paying a fine equal to 25% more than the value of the property taken, which would go into the state's general fund – and the other target had received an action-oriented punishment – paying

restitution to the victim of the same amount. Each punishment involved depriving the target of his own resources since he had to pay more than he had taken, and the amount of money was the same across conditions; the only difference was whether the money went to the state or the victim. We paired each vignette with pictures of two men from the Chicago Faces Database (Ma et al., 2015), which were matched on age, race, and a range of physical properties, as well as on ratings of traits like how angry, threatening, and trustworthy the faces appeared. To more fully control the effect of face perception on punishment judgments (Wilson & Rule, 2015), we also counterbalanced which face was associated with which punishment and which pairs were associated with which vignettes across participants. Three of the image pairs depicted White men and three depicted Black men, so that we could evaluate whether participants' inferences differed depending on the race of the targets.

Below is an example of the fraud vignette, with the images from the Chicago Faces database omitted:

> [Image of face omitted]
>
> **Person A**
>
> [Image of face omitted]
>
> **Person B**
>
> Each of these two people set up a fundraising website that he said was to help people impacted by a recent natural disaster, promoted the website on social media, but then kept all the money for himself.
>
> Based on the social media posts, police were able to arrest each person. Their punishments after conviction were the following:
>
> **Person A** was ordered to pay the state a fine equal to 25% more than the total money he received through his website, which went into the state's general fund. Person A paid the state.
>
> **Person B** was ordered to pay the victims back what they donated plus 25%, which

compensated the victims for the stolen money and the inconvenience. Person B paid the victims.

Participants then responded to three items asking about future behavior in counterbalanced order:

- One of these people committed another similar crime in the future. Which person do you think it was? (reverse-scored)
- One of these people turned his life around and never committed another crime. Which person do you think it was?
- Ten years in the future, one of these people won an award for services to his community. Which person do you think it was?

For each measure, participants made two judgments: a binary choice between the two targets and a rating of confidence in their choice on a five-point Likert scale, from "Not at all sure" to "Very sure." We converted these two responses into a score on a scale from -4.5 to 4.5 by multiplying two variables coded as follows: For the binary choice, we coded a choice of the target who received the action-oriented punishment to behave better in the future as 1 and a choice of the other target (who received a person-oriented punishment) as -1; for the confidence ratings, the lowest confident was coded as 0.5 and the highest as 4.5, with one-unit increments between the five levels.

Thus, each of the three future behavior items received a score where -4.5 reflected the highest confidence that the target who received a person-oriented punishment would behave better in the future and +4.5 reflected the high confidence that the target who received an action-oriented punishment would behave better in the future. Coding the confidence ratings from 0.5 to 4.5 ensured that, after multiplication, the distance between any two scale points was identical. Using integers would have created a situation where the difference between being "not at all" sure about choosing the target who received the person-oriented punishment and being "not at all" sure about choosing the target who received an action-oriented punishment would be a jump

from -1 to +1 – a difference twice as large as the difference between all other pairs of scale points.

Following this main measure of interest, each participant answered questions in counterbalanced order about possible covariates. Participants rated the seriousness of each crime using a 7-point Likert scale ranging from "not at all serious" to "very serious." They also rated the relative seriousness of each punishment on a 7-point Likert scale ranging from 1 (perceiving the person-oriented punishment as much more serious than the action-oriented punishment) to 7 (perceiving the action-oriented punishment as much more serious than the person-oriented punishment), with the midpoint indicating the perception that the two punishments were equally serious.

Finally, participants answered an attention check question and a series of demographic questions. The attention check question asked participants to recall and briefly describe one of the punishments appearing in any of the vignettes. Participants who answered the attention check correctly received $1.33; participants who did not received $0.10.

Transparency and Openness

The following applies to all studies reported in this paper and the Supplementary Materials: We report how we determined our sample size, all data exclusions, all manipulations, and all measures in this manuscript. All vignettes, measures, materials, data, and analysis code appear at the following link: https://osf.io/nxw24/?view_only=b92f21261c8e4804857bccd6099af96b.

The designs, planned data exclusions, and main analyses for each study are pre-registered. For all three studies, we pre-registered research questions rather than directional hypotheses. For Studies 1 and 2, the pre-registrations state what the dependent variable will

measure – i.e., which person between two punishment recipients is more likely to behave better in the future – and the coding method for the items that comprise that measure. The pre-registrations for these two studies do not include the exact number or wording of the items. However, we report all questions related to future behavior that participants answered and use all those questions to calculate the dependent variable for all studies. The three items used to calculate the dependent variable are the same across all three studies.

Results

We conducted four pre-registered analyses on the judgments about future behavior. We conducted all analyses in this section and throughout the paper using R v.4.2.2 (R Core Team, 2022).

First, we computed the dependent variable by averaging the responses (on the -4.5 to 4.5 scale described above) across all three items and all six vignettes (Cronbach's alpha = 0.95). We then regressed these mean responses on a constant, equivalent to a one-sample t-test comparing the mean responses to 0. The responses were significantly higher than 0 (M = 1.68, SD = 1.43, $t(210) = 17.09$, $p < .001$, $d = 1.18$). The positive mean indicates that participants expected better future behavior from the target who received an action-oriented rather than a person-oriented punishment.[6]

[6] An analysis of the ordinal data without averaging across vignettes and measures produced consistent results. We analyzed the ordinal responses on the 10-point scale ranging from the highest confidence the target receiving the person-oriented punishment would behave better in the future (0) to the highest confidence the target receiving the action-oriented punishment would behave better in the future (9) with a mixed model ordinal regression using a logit cumulative link function, with participant, vignette, and question treated as random factors in a crossed design, using the *ordinal* package in R (Christensen, 2022). Exponentiating the estimated ordinal thresholds to obtain the probability of classification into each ordinal category and converting to the -4.5 to 4.5 scale used in the main text by subtracting 4.5, the expected response in this model is 2.14 (95% CI: 1.92 to 2.33). In other words, based on this model, we would expect an average participant response of 2.14, which is slightly higher than the observed mean.

To assess whether the targets' race affected these judgments, we also conducted separate regressions comparing participants' mean responses to a constant for the three vignettes with Black targets and to the three vignettes with White targets. We found a significant difference from 0 for both sets of responses (Black: $M = 1.70$, $SD = 1.48$ $t(210) = 16.72$, $p < .001$, $d = 1.15$.; White: $M = 1.65$, $SD = 1.50$, $t(210) = 16.03$, $p < .001$, $d = 1.10$). A paired-sample t-test comparing the average responses to Black targets and the average response to White targets as a within-subject variable did not reveal a significant difference ($t(210) = -0.90$, $p = .37$, $d = -0.06$). These results appear in Figure 1.

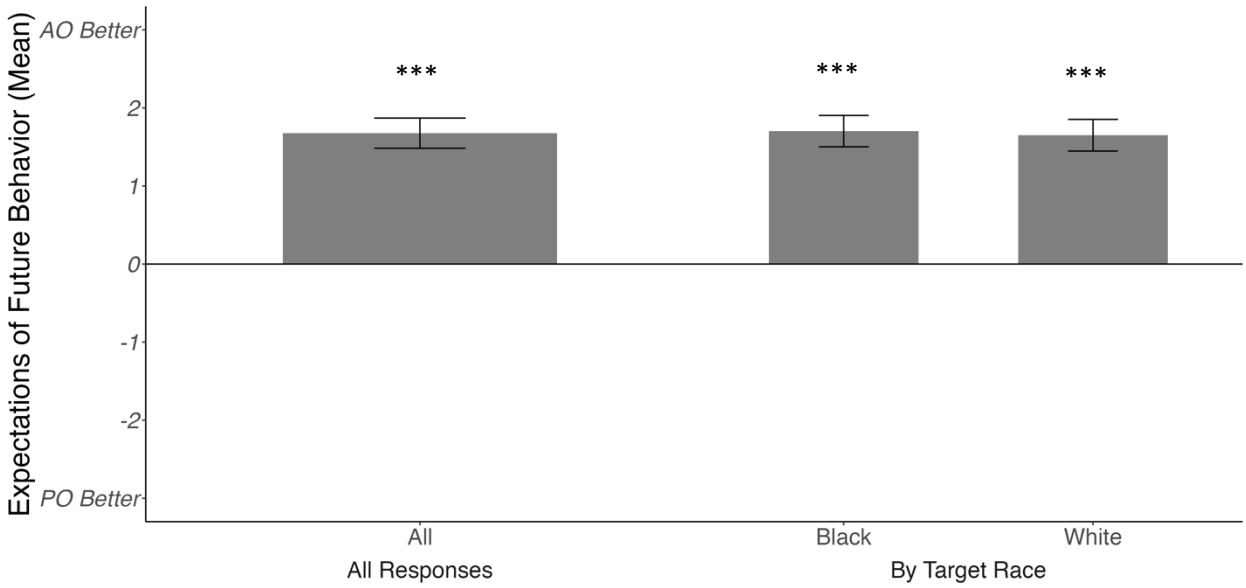


Figure 1. Average judgments of future behavior in Study 1 across all vignettes (All Responses) and across the three vignettes within each Target Race category (By Target Race). Positive values on the y-axis reflect judgments that the person who received an action-oriented punishment (AO) would behave better in the future; negative values reflect judgments that the person who received a person-oriented punishment (PO) would behave better in the future. Error bars show 95% confidence intervals.

In exploratory analyses, we also examined whether participants' expectations about future behavior may have arisen from a belief that one of the punishments was more severe than the other and therefore served as a greater deterrent. Regressing the comparative severity rating of the two punishment types on a constant showed that participants' severity ratings were significantly greater than the midpoint (4; $M = 0.52$, $SD = 1.09$, $t(210) = 6.93$, $p < .001$, $d = 0.48$), reflecting an average belief that the action-oriented punishment was more severe than the person-oriented punishment, even though the loss of money the targets experienced was identical in both cases.

However, this severity judgment did not drive the results of the main analysis. Regressing mean judgments of future behavior on a constant and mean-centered comparative severity judgments still showed an intercept significantly different from 0 – that is, an average prediction that the target who received the action-oriented punishment would behave better in the future ($\alpha = 1.68$, $t(209) = 17.497$, $p < .001$). This regression also showed a significant effect of comparative severity, such that an increase in the perceived severity of the action-oriented punishment predicted an increased certainty that the target who received the action-oriented punishment would behave better in the future ($\beta = 0.29$, $t(209) = 3.33$, $p = .001$). However, in this model, mean judgments remain significantly above 0 at both extremes of the severity scale. Recoding the severity variable so that the option originally coded as 1 ("Paying a fine to the state is much more severe") was coded as 0, the intercept remained significantly above 0 ($\alpha = 0.65$, $t(209) = 2.00$, $p = .047$), and the same was true when recoding the variable so the option originally coded as 7 ("Paying the victim is much more severe") was coded as 0 ($\alpha = 2.40$, $t(209) = 10.09$, $p < .001$). Restricting the sample to the 123 participants who thought the punishment

types were equally severe showed the same pattern ($M = 1.50$, $SD = 1.29$, $t(122) = 12.87$, $p < .001$, $d = 1.16$).

Discussion

Based on communicative theories of punishment (Bregant et al., 2016; Nahmias & Aharoni, 2017; Sarin et al., 2021; de Vel-Palumbo, 2023a), we asked whether participants expected punishment recipients to behave better in the future when they received an action-oriented punishment (which could communicate a didactic message that their behavior was bad) or a person-oriented punishment (which could communicate either an unclear message or, potentially, an ostracizing message that the recipients themselves were bad). Study 1 provided evidence that participants expected better future behavior from the target who received the action-oriented punishment.

Study 1 also addressed one possible alternative to the communicative explanation: That participants simply viewed one type of punishment as intrinsically more severe than the other, even though both required the target to pay equal amounts of money, and thought that the more severe punishment would create a stronger incentive not to engage in bad future behavior. Although a greater perceived severity of the action-oriented punishment did predict better ratings of future behavior for the target who received the action-oriented punishment, controlling for this severity effect did not eliminate the main effect or meaningfully reduce its size. Study 2 tested two additional alternative explanations for the results observed in Study 1.

Study 2

A key prediction of the communicative theory is that the person receiving the punishment must be aware of that punishment to receive its message. If the person receiving the punishment does not know she is being punished or is not aware of what type of punishment she is receiving,

then the communicative theory predicts that person's future behavior should depend on what she thinks occurred rather than what actually occurred (Sarin et al., 2021).

Two alternative explanations for the results obtained in Study 1, the *approval* and *private information* theories, make different predictions when a punishment recipient's beliefs deviate from reality. On the approval theory, participants in Study 1 rated the action-oriented punishment as more effective at improving future behavior because they approved of compensating the victim and were motivated to say that the punishment that led to compensation would also produce other desirable consequences. Thus, on this theory, participants who learned about two people who actually received different types of punishment – e.g., one paid money that ultimately compensated the victim and one paid money that ultimately went to the state – would still say they expected better future behavior from the person whose money went to the victim, even if that person believed that his money had gone to the state or the other person believed that his money had gone to the victim. In other words, the approval theory says that participants' judgments depended on where each person's money actually ended up, rather than the person's beliefs about whom he had paid.

On the private information theory, participants in Study 1 thought that imposing a person-oriented punishment reflected the judge's judgment, based on private information not known to participants, that the punishment recipient had a bad true self. On this theory, participants' judgments of future behavior depended on the punishment the judge intended to impose rather than the punishment the target believed he received. On the private information theory, therefore, participants who learned about two people upon whom judges intended to impose different types of punishment – e.g., to order one to pay money to the victim and the other to pay money to the state – would say they expected better future behavior from the person whom the judge intended

to order to pay the victim, even if that person believed that his money had gone to the state. Study 2 tested these alternative explanations for Study 1's results.

Methods

Participants

Based on Study 1, we expected a large effect size for the difference from zero in a planned condition that was structurally identical to Study 1 ($d \approx 1.1$). We expected the difference between the two groups used in Study 2 to be at least half that size ($d \approx 0.5$). Seeking a power of 0.8 and using a standard alpha level, we calculated in G*Power 3.1 that we needed a total sample size of 128 participants and recruited 291 adult participants on Amazon's Mechanical Turk to account for potential exclusions. In accordance with the pre-registration (https://aspredicted.org/MWD_S24), we excluded ninety-seven participants who completed the entire survey in less than one minute as measured using the Duration variable recorded by Qualtrics, did not complete the survey, or failed an attention check, leaving an analysis set of 194 participants (M_{age} = 41.35 years, SD_{age} = 12.54 years; 57% female, 42% male, the remaining participants did not answer this question).[7] Based on a sensitivity analysis, this sample size – allocated between the no-mistake condition (n=88) and the mistake condition (n=106), both described below – allowed us to detect an effect of size d = 0.41 or greater, given a desired power of 0.8 and a standard alpha level.

Participants self-identified their race in the following proportions: 80% White or European-American, 7% Asian or Asian-American, 8% Black or African-American, 5% multiracial, and 0.5% "option not listed." Additionally, 7% of participants self-identified as

[7] Most of these exclusions resulted from participants failing the attention check. The proportion of exclusions is consistent with recent studies on this platform (e.g., Murray et al., 2023).

Hispanic or Latino. The sample's highest level of education was distributed as follows: 6% high school or GED, 19% some college or university, 6% associate's degree, 47% bachelor's degree, 17% master's or professional degree, 4% PhD; the remaining participants did not answer this question.

Procedure

This study tested the communicative theory against the approval and private information theories by randomly assigning participants to one of two between-subject conditions, the *no-mistake* condition and the *mistake* condition, designed so that the communicative theory predicted a significant difference between conditions but the other two theories did not. In the no-mistake condition, participants received the same pictures and vignettes used in Study 1, with only slight modifications in how the money got to the ultimate recipient (described below). In the mistake condition, the vignettes included further modifications so that the punishments the two targets ultimately received were the same as in the no-mistake condition (i.e., one target's money actually went to the state and one target's money actually went to the victim(s)), the punishments the judge intended to impose on the two targets was also the same as in the no-mistake condition (i.e., one judge intended to impose a person-oriented punishment and one judge intended to impose an action-oriented punishment), but one of the two targets had a mistaken belief about which type of punishment he received (i.e., he thought his payment was going to the victim, but it actually went to the state, or vice versa). Thus, in the mistake condition, either both targets believed they received a person-oriented punishment or both targets believed they received an action-oriented punishment.

The materials described the mistake in the mistake condition as follows: The judge sentencing one target intended to impose a fine and the judge sentencing the other target

intended to impose restitution. One judge made a mistake at sentencing and incorrectly announced a different type of punishment than he had intended, so the two targets received the same type of punishment. After sentencing, the mistaken judge caught the mistake and corrected the judgment, but the target did not receive the notice from the court about the correction. Both targets paid their money to the court and the court clerk ensured it went to the right recipient.[8] Thus, when the mistaken target paid his money to the court, he believed the money would go to one recipient (either the court itself or the victim(s)), but the clerk sent it to the other recipient, based on the corrected judgment.

As in Study 1, each participant received the six vignettes in counterbalanced order and each participant randomly received half the vignettes paired with pictures of Black men and half the vignettes paired with pictures of White men. Participants answered questions using the same measures of future behavior used in Study 1, in counterbalanced order. Their responses were coded from -4.5 to 4.5 in the same way described in the Study 1 procedure. In the mistake condition, each participant received three vignettes in which the person who received the person-oriented punishment was mistaken and three vignettes in which the person who received the action-oriented punishment was mistaken; we counterbalanced which vignettes were in each group.

Below is an example of the fraud vignette in the mistake condition, with underlining indicating text that differed from Study 1:

[8] Study 1 did not explain the mechanics of how the victim received the money for the action-oriented punishment. This additional information appeared in both conditions of Study 2. As the results show, the no-mistake condition replicated the findings in Study 1, suggesting that the result did not arise from participants assuming that the target who received the action-oriented punishment repaid their victim in person.

> Each of these two people set up a fundraising website that he said was to help people impacted by a recent natural disaster, promoted the website on social media, but then kept all the money for himself.
>
> Based on the social media posts, police were able to arrest each person. Each person was convicted in court and sentenced by a judge. The judges imposed the following sentences:
>
> Person A was ordered to pay the victims back what they donated plus 25%, to compensate the victims for the stolen money and the inconvenience. This sentence was exactly what the judge intended to order.
>
> Person B was ordered to pay the victims back what they donated plus 25%, to compensate the victims for the stolen money and the inconvenience. The judge had intended to order the defendant to pay this money to the court as a fine, but the judge made an error when imposing the sentence. The judge later corrected the judgment but Person B did not receive the notice from the court and wasn't aware of the correction.
>
> Both Person A and Person B paid the money to the court, so the court could forward it to the victims. The clerk of the court that received Person B's payment noticed that Person B's judgment had been updated and kept the money at the court as payment of the fine.

As in Study 1, after answering questions about all six vignettes, participants received a question about the severity of the two types of punishment, the attention checks, and demographics questions. Due to an increase in low-quality responses on Amazon's Mechanical Turk around the time we conducted Study 2 (Chmielewski & Kucker, 2020), this experiment used more stringent attention checks than the first study. In addition to the question asking participants for a detail from the stories that anyone paying attention should have remembered (a brief description of one of the punishments described), the attention checks included three true-or-false questions modeled on the types of questions that large language models have difficulty answering (e.g., addition in word problems, common knowledge). Participants who answered the attention checks correctly received $1.33; participants who did not received $0.10.

Results

We conducted three pre-registered analyses on the participants' judgments about the targets' future behavior. First, we replicated the analysis from Study 1 within each mistake condition and across the full sample. Averaging across all three items and all six vignettes (Cronbach's alpha = 0.89), we conducted two separate linear regressions of these mean responses on a constant, equivalent to a one-sample t-test. Participants' mean predictions of future behavior were significantly greater than 0 aggregated across the whole analysis sample (M = 0.88, SD = 1.16, $t(193) = 10.54$, $p < .001$, $d = 0.76$), within the no-mistake condition (M = 1.53, SD = 1.23, $t(87) = 11.7$, $p < .001$, $d = 1.25$), and within the mistake condition (M = 0.34, SD = 0.77, $t(105) = 4.57$, $p < .001$, $d = 0.44$). These results replicated the results of Study 1 and showed that, on average, participants expected targets to behave better in the future after receiving action-oriented rather than person-oriented punishments. Second, we compared participant responses between the two mistake conditions. We conducted a linear regression of these mean responses on an indicator variable for the mistake condition (dummy-coded so that 1 = mistake condition and 0 = no-mistake condition), equivalent to an independent samples t-test. The difference between conditions was significant and large in size (β = -1.189, $t(192) = -8.23$, $p < .001$, d = 1.19). All the results appear in Figure 2.

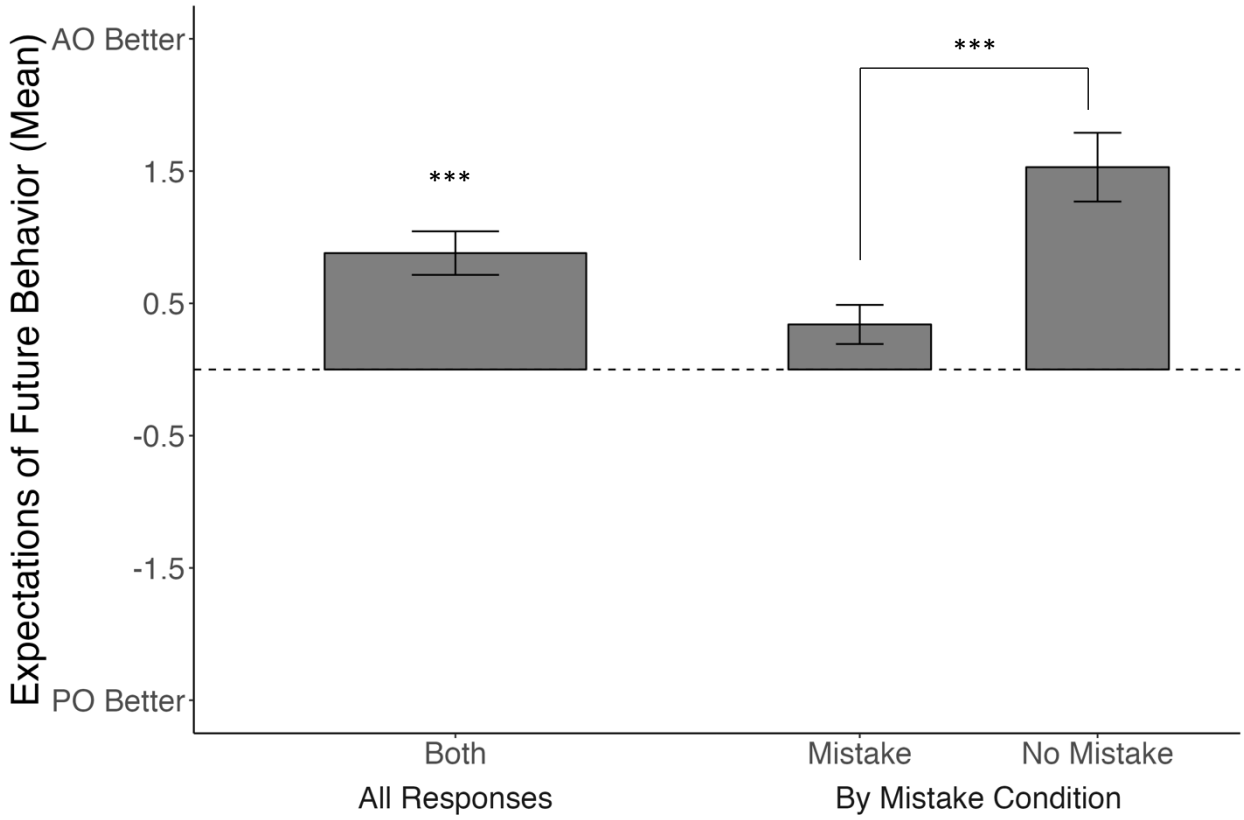


Figure 2. Average judgments of future behavior in Study 2 across all vignettes and participants (All Responses) and separated by mistake condition (By Mistake Condition). Positive values on the y-axis reflect judgments that the person who actually received an action-oriented punishment (AO) would behave better in the future; negative values reflect judgments that the person who actually received a person-oriented punishment (PO) would behave better in the future. Error bars show 95% confidence intervals.

As in Study 1, we did not find a significant effect of the race of the targets on the relative predictions of future behavior. A 2 (Judgment Type: mistake vs. no mistake) x 2 (Target Race: Black vs. White) ANOVA with repeated measures on the second factor showed a significant main effect of Judgment Type ($F(1, 193) = 67.68$, $p < .001$, partial $\eta^2 = .26$), no significant effect of Target Race ($F(1, 194) = 2.51$, $p = .115$, partial $\eta^2 = .01$), and no significant interaction ($F(1,$

194) = 1.86, p = .175, partial η^2 < .001). Mean judgments did not differ for Black and White targets in the no-mistake condition (both Ms = 1.53, $SD_{\text{Black target}}$ = 1.40, $SD_{\text{White target}}$ = 1.34). The marginal effect of race arose entirely from a marginally significant difference in predictions of future behavior in the mistake condition ($M_{\text{Black target}}$ = 0.20, $M_{\text{White target}}$ = 0.48, $SD_{\text{Black target}}$ = 1.10, $SD_{\text{White target}}$ = 1.06, t(105) = 1.91, p = .058, d = 0.19). Responses in the mistake condition were only marginally above zero for Black targets (t(105) = 1.88, p = .063, d = 0.18) but were significantly above zero for White targets (t(105) = 4.68, p < .001, d = 0.45).

As in Study 1, we also conducted exploratory analyses using participants' judgments of the relative severity of action-oriented and person-oriented punishments to assess whether those judgements explained participants' predictions regarding targets' future behaviors. As in Study 1, a regression of participants' comparative severity judgments on a constant revealed that participants viewed the action-oriented punishment as more severe than the person-oriented punishment, as reflected in mean severity ratings that were significantly higher than the midpoint (4; M = 4.38, SD = 1.11, t(195) = 4.85, p < .001). Regressing the judgments of future behavior, averaged across measures and vignettes, on the mistake condition dummy variable and the centered comparative severity judgment produced a model with a significant intercept (α = 1.50, t(191) = 14.38, p < .001), a significant effect of mistake condition (β = -1.14, t(191) = -8.05, p < .001), and a significant effect of centered comparative severity (β = 0.21, t(191) = 3.23, p = .001). As in Study 1, mean judgments remained significantly above 0 at both extremes of the severity scale. Recoding the severity variable so that the option originally coded as 1 ("Paying a fine to the state is much more severe") was coded as 0, the intercept remained significantly above 0 (α = 0.80, t(191) = 3.27, p = .001), and the same was true when recoding the variable so the option originally coded as 7 ("Paying the victim is much more severe") was coded as 0 (α =

2.04, $t(191) = 10.76$, $p < .001$). As in Study 1, participants appeared to factor comparative severity into their judgments of future behavior, but perceived punishment severity did not explain most of the variance in comparative punishment judgments.

Finally, to investigate whether concerns about procedural justice may have played a role in these results, we also disaggregated mistake-condition trials in which there was a mistake in the sentence of the target who actually received the person-oriented punishment ($M = 0.69$, $SD = 1.45$) and those in which there was a mistake in the sentence of the target who actually received the action-oriented punishment ($M = -0.01$, $SD = 1.44$). A paired samples t-test showed a significant difference between these two groups of vignettes, $t(105) = 2.93$, $p = .004$, $d = 0.28$. The mean of the no-mistake condition was significantly higher than both the former, $t(191.91) = 4.38$, $p < .001$, $d = 0.62$, and the latter, $t(191.87) = 8.02$, $p < .001$, $d = 1.14$, means.

Discussion

Study 2 provided evidence that the communicative meaning of a punishment largely drove participants' judgments that the action-oriented punishment would produce better future behavior than the person-oriented punishment. In both conditions, the targets knew how much money they were losing. When both targets knew who had received their payments, participants predicted that the target who received the action-oriented punishment would behave better in the future, as in Study 1. However, when both targets believed (one mistakenly) that their payments were going to the same type of recipient – either the state or the victim – participants were significantly less certain that actually receiving the action-oriented punishment would lead to better future behavior. This difference indicates that the targets' beliefs about who received their payments played an important role in participants' judgments of the effectiveness of different types of punishments.

The communicative theory is the best explanation for the importance of the targets' beliefs to participants' judgments. The approval theory depends on where the targets' money actually ended up, and the private information theory depends on where the judges intended the targets' money to go. In this experiment, neither the ultimate destination of the money nor the judges' intentions varied between conditions – in both conditions, one target's money ultimately ended up with the victim and the other target's money ultimately ended up with the state, and the judges each intended that outcome. Of the theories under consideration, only the communicative theory predicted the difference between conditions observed in this study.

Finally, an exploratory analysis addressed another potential explanation for the mean differences observed in this study. Prior research has shown that people show less motivation to behave prosocially after punishment imposed in a procedurally unjust manner (de Vel-Palumbo et al., 2023a; Maguire et al., 2017). This finding raises a concern: Participants may have perceived the mistake condition to involve a procedural injustice – e.g., a judge paying so little attention during sentencing that they erroneously imposed the wrong punishment. If that were the case, mean responses could have been near 0 in the mistake condition because participants viewed the procedural injustice as a stronger influence than the punishment type and always predicted that the target whose sentence did not involve a mistake would behave better in the future. On this view, the average in the version of the mistake condition where the mistake affected the target who actually received the person-oriented punishment would be positive (i.e., better future behavior by the target who actually received the action-oriented punishment) and the average in the version where the mistake affected the target who actually received the action-oriented punishment would be negative (i.e., better future behavior by the target who actually received the person-oriented), yielding a near-zero mean in the mistake condition as a whole.

The final exploratory analysis is inconsistent with this proposal. Although participants on average predicted better future behavior from the target who received the action-oriented punishment when the mistake affected the target receiving the person-oriented punishment, the reverse pattern did not appear when the mistake affected the target who received the action-oriented punishment. Likewise, mean responses when the mistake affected the target who received the person-oriented punishment were significantly lower than mean responses in the no-mistake condition, rendering it unlikely that the procedural injustice had a stronger effect on participants than the punishment type.

Study 3

According to the communicative theory, participants' predictions about the future behavior of a punishment recipient depend on that person's beliefs about the punishment they received, because those beliefs determine what the punishment communicated. Action-oriented punishment, which forces a recipient to act out normative behavior, teaches a more effective lesson than person-oriented punishment, which merely harms the recipient. Participants therefore expect a person who believes he has received action-oriented punishment to behave better in the future than a person who believes he has received a person-oriented punishment. For the same reason, participants have no basis for deciding which of two people with the same beliefs about the type of punishment they received will behave better in the future.

This theory therefore makes two predictions beyond those tested in Study 2. Recall that Study 2 compared a scenario in which there were no mistakes in the punishments imposed on either target (i.e., a scenario with zero mistakes in sentences across the two targets) to a scenario in which there was a mistake in the punishment of one of the two targets (i.e., a scenario with one mistake in a sentence across the two targets). The communicative theory predicted that

participants would be closer to indifferent about which target would behave better in the future (i.e. responses would be much closer to 0) in the one-mistake scenario, since there the two targets had the same beliefs about the punishments they received, and predictions about future behavior depend on those beliefs alone according to this theory.

Now, consider what the communicative theory predicts if *both* targets are mistaken about the punishments they have received (i.e., a scenario with two mistakes in the sentences, one for each of the two targets). In that scenario, the target who actually received a person-oriented punishment would believe he received an action-oriented punishment and vice versa. Thus, in this two-mistake scenario, the communicative theory predicts that participants will expect better future behavior from the target who actually received the person-oriented punishment because he *believed* he received the action-oriented punishment, while the other target, who actually received the action-oriented punishment, believed he received the person-oriented punishment. That is, the theory predicts that participants' judgments will show a reversal of the zero-mistake scenario, by predicting better future behavior from the target who actually received the person-oriented punishment rather than the target who actually received the action-oriented punishment.

Put another way, the number of mistakes – zero, one, or two – should have a negative (statistical) effect on expectations of future behavior, as measured on the scale used in the present studies. Zero mistakes should elicit a positive mean (i.e., a prediction that the person who actually received, and uniquely believed he received, the action-oriented punishment will behave better in the future), one mistake should elicit a mean around 0 (i.e., no clear prediction about which of two people with the same beliefs about the punishments they received will behave better in the future), and two mistakes should elicit a negative mean (i.e., a prediction that the

person who actually received the person-oriented punishment, but uniquely believed he received the action-oriented punishment, will behave better in the future).

The communicative theory also predicts that judgments about learning underlie this pattern of judgments. Participants think action-oriented punishments communicates more about the reasons a transgression was wrong and what behavior is normative than person-oriented punishments. Thus, if the communicative theory is correct, then judgments about which target learned more about how to behave should mediate this effect of mistakes on behavioral expectations.

This study tests these two predictions. In doing so, it can also provide evidence regarding an alternative explanation for the effect observed in Study 2 – namely, that participants found the mistake condition confusing and therefore chose randomly between the two targets when making behavioral predictions, leading to the near-zero mean. This *confusion* account would predict no significant difference between the one-mistake and two-mistake conditions. Even if participants were more confused in the two-mistake condition and thus more likely to choose randomly, this account would, at most, predict that the mean response in the two-mistake condition would be even more likely to average to 0 than those in the one-mistake condition. Similarly, this study can provide further evidence about the procedural injustice theory raised (and not supported) in the discussion section of Study 2. If participants always expected worse future behavior from targets who experienced procedural injustice and viewed the mistake as a procedural injustice, then they should have been indifferent between the two targets in the two-mistake condition and provided responses with a near-zero mean in that condition.

Methods

Participants

Study 3 used a 2 (Mistake for Person-Oriented Punishment Recipient: yes vs. no) x 2 (Mistake for Action-Oriented Punishment Recipient: yes vs. no) design. Based on a simulation using the means and standard deviations observed in Study 2 (see Supplementary Material), we expected large main effects (partial $\eta^2 \approx .14$) and a small interaction effect (partial $\eta^2 < .03$). We expected the main effects to correspond to lower means when a mistake was present as opposed to absent for either predictor. An interaction effect that increased the mean substantially when there were mistakes in both sentences would be inconsistent with the communicative theory. Therefore, to establish the pattern of results relevant to our predictions, we only needed sufficient power to detect an interaction effect greater than this size. Seeking a power of 0.8 and using a standard alpha level, we used G*Power 3.1 to calculate that we needed a total sample size of 256 participants and recruited 301 adult participants on Prolific to account for potential exclusions. In accordance with the pre-registration (https://aspredicted.org/2FC_V3L), we excluded eight participants who failed an attention check, did not agree that they were human, agreed that they were a large language model, or did not complete the study,[9] leaving an analysis set of 293 participants (M_{age} = 37.50 years, SD_{age} = 11.64 years; 52% female, 46% male, 2% non-binary). These participants were allocated between the four conditions as follows: no PO mistake and no AO mistake, n = 85; PO mistake and no AO mistake, n = 65; no PO mistake and AO mistake, n = 79; PO mistake and AO mistake, n = 64. Based on a sensitivity analysis conducted using G*Power 3.1, this sample size allowed us to detect effect sizes of partial $\eta^2 \approx .026$ or greater with a power of 0.8, using a standard alpha level.

[9] The other exclusion criteria we planned to apply – completing the entire study in less than 1 minute or less than three standard deviations below the mean completion time – did not exclude any participants.

Participants self-identified their race in the following proportions: 60% White or European-American, 16% Black or African-American, 14% Asian or Asian-American, 6% multiracial, 2% Native American or Pacific Islander, and 2% "option not listed." Additionally, 10% of participants self-identified as Hispanic or Latino. The sample's highest level of education was distributed as follows: 1% less than high school, 12% high school or GED, 25% some college or university, 10% associate's degree, 38% bachelor's degree, 9% master's degree, 4% professional or doctoral degree; the remaining participants did not answer this question.

Procedures

We randomly assigned each participant to one of four between-subjects conditions in a 2 (Mistake for Person-Oriented Punishment Recipient: yes vs. no) x 2 (Mistake for Action-Oriented Punishment Recipient: yes vs. no) design. As in Studies 1 and 2, participants received six vignettes in counterbalanced order describing two people (Person A and Person B) who had separately committed similar property crimes (larceny, robbery, armed robbery, burglary, fraud, and identity theft) and were convicted of those crimes in court proceedings. Participants read that one person ultimately received an action-oriented punishment and paid an amount equal to 25% more than the value of the stolen property to the victim(s), while the other person ultimately received a person-oriented punishment and paid an amount equal to 25% more than the value of the stolen property to the state. In the conditions where one or more mistakes occurred, participants read a description of how the mistake occurred that was identical to the description in Study 2. The vignettes were presented along with a randomly selected pair of pictures, labeled "Person A" and "Person B." Each participant randomly received half the vignettes with pictures of two Black men and half the vignettes with pictures of two White men.

After reading the vignette, participants received the learning from punishment measure (the proposed mediator) and the future behavior measure (the dependent variable) on separate pages. The learning from punishment measure contained three forced-choice questions asking participants to choose the target that learned a lesson, received better guidance about how to behave, and really understands why what he did was wrong.[10] Participants then rated their confidence in this judgment on a five-point Likert scale. We converted these binary choices and confidence ratings into responses on a scale from -4.5 (strong confidence that the target receiving person-oriented punishment learned more) to 4.5 (strong confidence that the target receiving action-oriented punishment learned more) in the same way described for Study 1. The future behavior measure contained the same questions, coded the same way, as in Studies 1 and 2.

At the top of the pages containing each of these measures, participants read a short overview of the punishments received by each person. For example, if Person A had received the action-oriented punishment, Person B received the person-oriented punishment, and there was a mistake only in Person B's sentencing, participants read the text below before making judgments:

> Person A paid the money to the court, believing that the court would forward the money to the victim. The court forwarded the money to the victim.
>
> Person B paid the money to the court, believing that the court would forward the money to the victim. Based on the updated judgment, the court kept the money as payment of the fine instead.

[10] The pre-registration describes these three items and how they are used to compute the proposed mediator.

As in Study 1, after answering questions about all six vignettes, participants received a question about the severity of the two types of punishment, a memory question, and demographics questions. The memory question asked participants for a detail from the stories that anyone paying attention should have remembered (a brief description of one of the punishments described). In all analyses below, we include responses from participants who failed the memory question, since that was not a pre-registered exclusion criterion for this study, but the pattern of results is identical if we exclude them.

Results

We conducted five pre-registered analyses on the participants' judgments about the targets' future behavior. First, we conducted a 2 (mistake in sentence of the target actually receiving person-oriented punishment: yes vs. no) x 2 (mistake in sentence of the target actually receiving action-oriented punishment: yes vs. no) ANOVA with the future behavior measure as the dependent variable. There was a main effect of the mistake in the person-oriented punishment, $F(1, 289) = 6.97$, $p = .009$, partial $\eta^2 = .12$, a main effect of the mistake in the action-oriented punishment, $F(1, 289) = 41.74$, $p < .001$, partial $\eta^2 = .29$, and an interaction effect $F(1, 289) = 5.80$, $p = .017$, partial $\eta^2 = .02$. Mean future behavior ratings for the four groups appear in Figure 3A. These effect sizes for the main effects and the interaction were close to the effect sizes predicted by the simulation (see Supplementary Materials).

Second, we used independent samples *t*-tests to compare the means of these four groups on the future behavior measure. Participants in the condition where neither target had a mistake in their sentence ($M = 1.83$, $SD = 1.42$) were significantly more confident the target receiving the action-oriented punishment would behave better in the future than participants in the condition where only the target receiving the person-oriented punishment had a mistake in his sentence (*M*

= 1.21, *SD* = 1.50), *t*(148) = 2.58, *p* = .011, *d* = 0.43, and participants in the condition where only the target receiving the action-oriented punishment had a mistake in his sentence (*M* = 0.39, *SD* = 1.45), *t*(162) = 6.40, *p* < .001, *d* = 1.00). Likewise, participants in the condition where only the target receiving the person-oriented punishment had a mistake in his sentence were significantly more confident the target receiving the action-oriented punishment would behave better in the future than participants in the condition where both targets had mistakes in their sentences (*M* = -1.04, *SD* = 1.30), *t*(127) = 9.06, *p* < .001, *d* = 1.60, as were participants in the condition where only the target receiving the action-oriented punishment had a mistake in his sentence, *t*(141) = 6.11, *p* < .001, *d* = 1.03. Unexpectedly, the two one-mistake conditions differed in means, with participants in the condition where only the target receiving the person-oriented punishment had a mistake in his sentence demonstrating significantly more confidence that the target receiving action-oriented punishment would behave better in the future than participants in the condition where only the target receiving the action-oriented punishment had a mistake in his sentence, *t*(142) = 3.31, *p* = .001, *d* = 0.55.

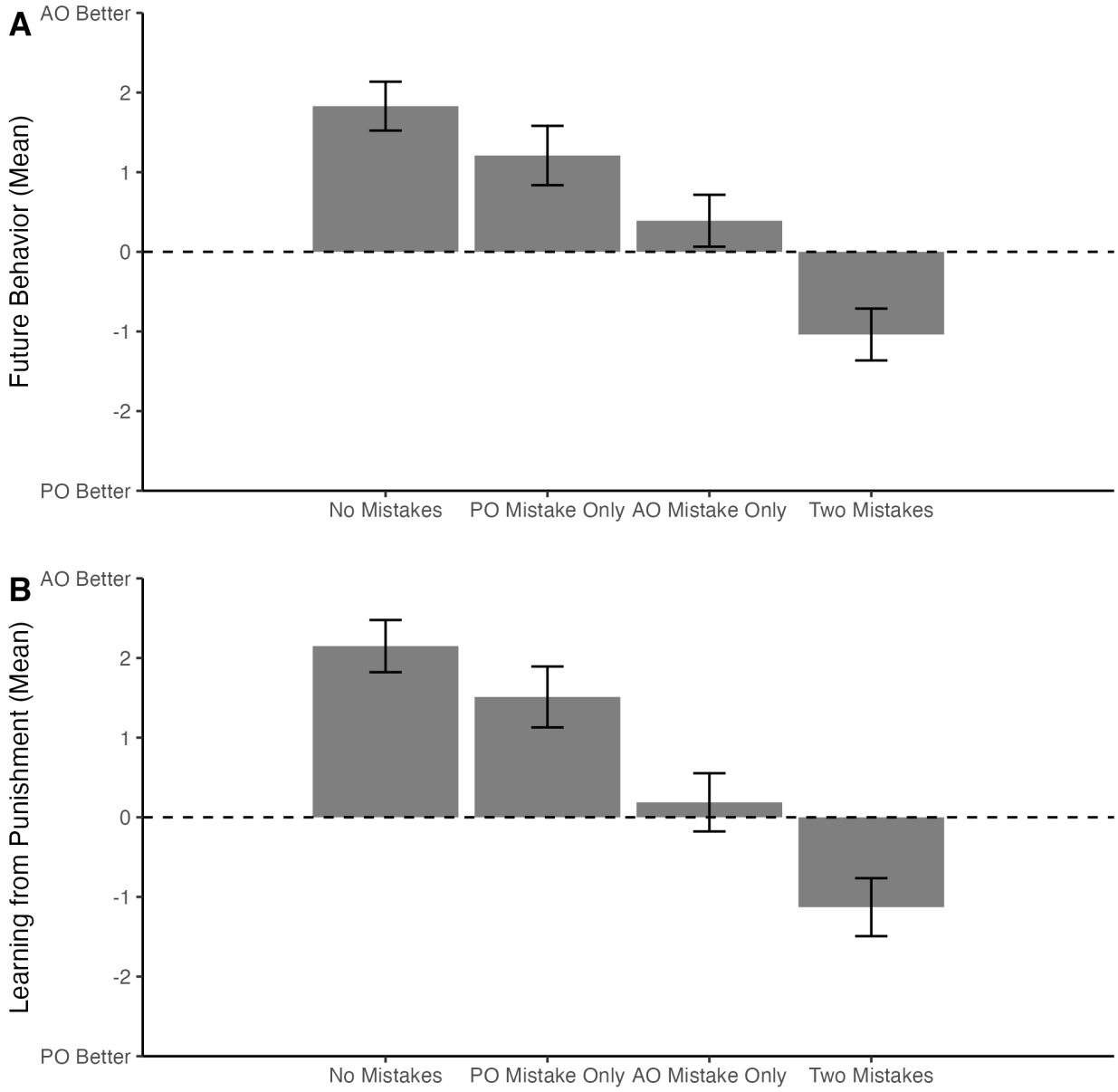
Third, the means on the future behavior measure for all four groups differed significantly from zero. Participants were significantly more confident that the target who actually received the action-oriented punishment would behave better in the future in the condition with no mistakes, *t*(84) = 11.85, *p* < .001, *d* = 1.29, the condition with just a mistake in the person-oriented sentence, *t*(64) = 6.48, *p* < .001, *d* = 0.80, and the condition with just a mistake in the action-oriented sentence, *t*(78) = 2.38, *p* = .020, *d* = 0.27. Participants were significantly more confident that the target who actually received the person-oriented punishment (but believed he received the action-oriented punishment) would behave better in the future in the condition with two mistakes, *t*(63) = -6.37, *p* < .001, *d* = -0.80.

Fourth, we defined a new independent variable that equaled the number of mistakes in each condition. This independent variable had three levels: 0 mistakes (no mistake at sentencing for either target), 1 mistake (a mistake at sentence for the target who actually received the action-oriented punishment or the target who actually received the person-oriented punishment, but not both), and 2 mistakes (mistakes at sentencing for both targets). A one-way ANOVA showed a significant effect of number of mistakes on the future behavior measure, $F(1, 291) = 137.14$, $p < .001$, partial $\eta^2 = .32$. Independent samples t-tests showed a significant difference in means between the 0 mistake condition ($M = 1.83$, $SD = 1.42$) and the 1 mistake condition ($M = 0.76$, $SD = 1.53$), $t(227) = 5.25$, $p < .001$, $d = 0.72$, as well as between the 1 mistake condition and the 2 mistake condition ($M = -1.04$, $SD = 1.30$), $t(206) = 8.18$, $p < .001$, $d = 1.23$, on the future behavior measure. Finally, the 1 mistake condition mean on this measure differed significantly from 0, $t(143) = 5.97$, $p < .001$, $d = 0.50$, as did the means in the 0 mistake condition and the 2 mistake condition, both reported above.

Fifth, we found a similar pattern for the learning from punishment measure. A one-way ANOVA showed a significant effect of number of mistakes on this measure, $F(1, 291) = 148.99$, $p < .001$, partial $\eta^2 = .34$. Independent samples t-tests showed a significant difference in means between the 0 mistake condition ($M = 2.15$, $SD = 1.52$) and the 1 mistake condition ($M = 0.79$, $SD = 1.72$), $t(227) = 6.05$, $p < .001$, $d = 0.83$, as well as between the 1 mistake condition and the 2 mistake condition ($M = -1.13$, $SD = 1.46$), $t(206) = 7.75$, $p < .001$, $d = 1.16$, on the learning from punishment measure. Finally, the means in all three conditions differed significantly from 0 (0 mistake condition mean: $t(84) = 13.04$, $p < .001$, $d = 1.41$; 1 mistake condition mean: $t(143) = 5.48$, $p < .001$, $d = 0.46$; 2 mistake condition mean, $t(63) = -6.19$, $p < .001$, $d = -0.77$). Figure 3B depicts the means for the learning from punishment measure.

Figure 3

Mean predictions of future behavior and learning from punishment judgments in Study 3.



Note. (A) Mean predictions of future behavior. (B) Meaning learning from punishment judgments. Positive values on the y-axis reflect confidence that the person who *actually* received the action-oriented (AO) punishment would behave better in the future/learn more than the

person who *actually* received the person-oriented (PO) punishment, while negative values reflect the opposite. All error bars are 95% confidence intervals.

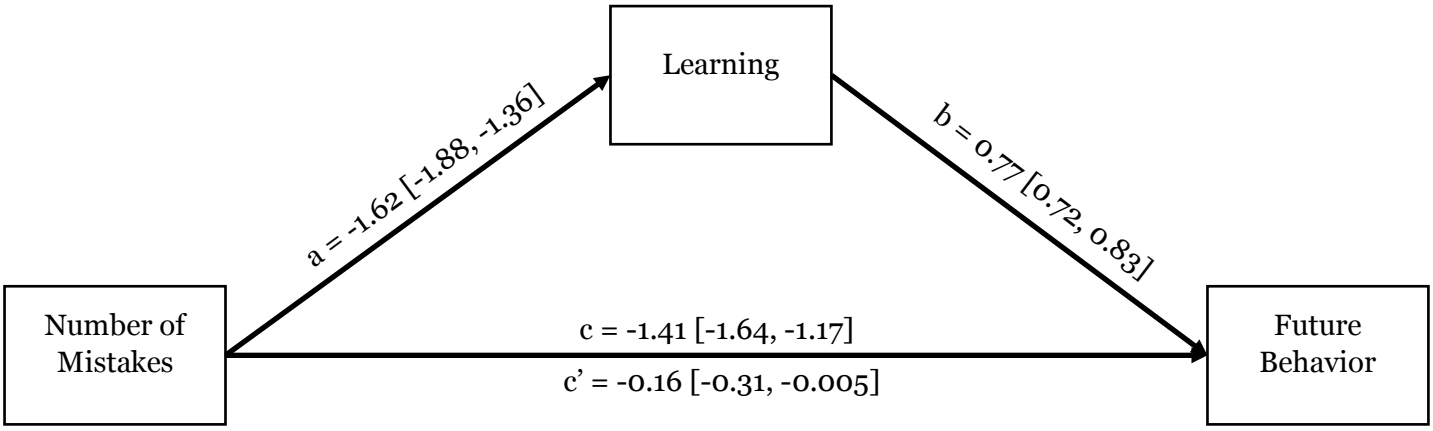
Due to the way the 0, 1, and 2 mistake conditions were constructed, the communicative theory predicts a negative relationship between number of mistakes and future behavior – i.e., predicting positive responses in the 0 mistake condition, reflecting a belief the person receiving the action-oriented punishment would behave better in the future; predicting near-0 responses in the 1 mistake condition, reflecting indifference between people with the same beliefs about the punishments they received; and predicting negative responses in the 2 mistake condition, reflecting a belief the person actually receiving the person-oriented punishment (but believing he received the action-oriented punishment) would behave better in the future. The communicative theory predicts a similar relationship between number of mistakes and learning from punishment, as well as a mediation of the former relationship by the latter.

To test whether perceived learning from punishment mediated the effect of number of mistakes on predictions of future behavior, we conducted a mediation analysis using the PROCESS macro v.4.3 (Hayes, 2017) for R with 5,000 bootstrapped samples. In this analysis, the experimental manipulation (Number of Mistakes) served as the predictor. We entered judgments about learning from punishment (Learning) as the mediator and judgments about future behavior (Future Behavior) as the outcome variable, rather than vice versa, because that order reflects both the order in which learning and future behavior would occur in the world (i.e., the learning from punishment judgments are about a punishment that occurs *before* the future behavior) and the order in which the participants answered these questions. That is, participants could not adjust their learning judgments after they had proceeded to the future behavior

judgments, and the learning judgments concerned what the punishment taught the targets, while the future behavior judgments concerned how the targets would behave after receiving the punishment (and learning whatever lesson it taught). This analysis found a significant mediation as reflected in Figure 4.

Figure 4

Mediation diagram for Study 3.



Note. Numbers in brackets are 95% confidence intervals calculated using 5,000 bootstrapped samples. The *c* path reflects the direct effect of number of mistakes on judgments about future behavior when the mediator is not included in the model. The *c'* path reflects the direct effect after including the mediator. The total indirect effect (*ab*) is -1.25 [-1.48, -1.03].

As in Studies 1 and 2, an exploratory analysis showed no effect of the targets' race. Paired-sample t-tests did not show any significant effect of the targets' race on any of the future behavior means (all *p*s > 0.51) or any of the learning from punishment means (all *p*s > 0.53).

As in Studies 1 and 2, we also conducted exploratory analyses using participants' judgments of the relative severity of action-oriented and person-oriented punishments to assess whether those judgements explained participants' predictions regarding the targets' future

behaviors. As in Study 1, a regression of participants' comparative severity judgments on a constant revealed that participants viewed the action-oriented punishment as more severe than the person-oriented punishment (4; $M = 4.37$, $SD = 1.17$, $t(292) = 5.33$, $p < .001$). Regressing the judgments of future behavior, averaged across measures and vignettes, on the two mistake condition dummy variables, their interaction, and the centered comparative severity judgment produces a model with a significant intercept ($\alpha = 1.85$, $t(288) = 12.06$, $p < .001$), a significant effect of a mistake in the sentence of the target receiving the person-oriented punishment ($\beta = -0.63$, $t(288) = -2.73$, $p = .007$), a significant effect of a mistake in the sentence of the target receiving the action-oriented punishment ($\beta = -1.46$, $t(288) = -6.63$, $p < .001$), a significant interaction ($\beta = -0.81$, $t(288) = -2.45$, $p = .015$), and a significant effect of centered comparative severity ($\beta = 0.18$, $t(288) = 2.59$, $p = .010$). As in Study 1, mean judgments remain significantly above 0 at both extremes of the severity scale. Recoding the severity variable so that the option originally coded as 1 ("Paying a fine to the state is much more severe") is coded as 0, the intercept remains significantly above 0 ($\alpha = 1.23$, $t(288) = 4.47$, $p < .001$), and the same is true when recoding the variable so the option originally coded as 7 ("Paying the victim is much more severe") is coded as 0 ($\alpha = 2.33$, $t(288) = 9.44$, $p < .001$). As in Studies 1 and 2, participants appeared to factor comparative severity into their judgments of future behavior, but perceived punishment severity did not explain most of the variance in comparative punishment judgments.

Discussion

Study 3 tested two additional predictions of the communicative model. First, Study 3 provided further evidence of the centrality of a punished person's beliefs about the punishment he received to participants' predictions about his future behavior. When both targets were mistaken about the punishments they would receive, participants continued to predict that the

target who believed he received the action-oriented punishment (but actually received the person-oriented punishment) would behave better in the future than the target who believed he received the person-oriented punishment (but actually received the action-oriented punishment).

Second, Study 3 provided additional evidence that these beliefs affect the targets' future behavior, in the participants' view, by teaching the targets how to behave normatively (or not). Judgments about what the targets learned from their punishment significantly mediated the effect of the number of mistakes on the predictions of future behavior. In other words, the mistake scenarios changed participants' judgments about future behavior largely by changing their judgments about which target would learn more from punishment.

In addition to supporting the predictions of the communicative model, the results of this study did not support the predictions of the confusion account. Mean future behavior ratings were significantly below 0 in the two-mistake condition, not indistinguishable from 0 as this account predicts. For the same reason, the results of this study suggest that inferences about the effect of procedural justice on learning and/or future behavior are not the primary drivers of the effect.

The results of this study also replicate the unexpected exploratory finding in Study 2 showing a significant difference between the two one-mistake conditions. When both targets believed they had received the action-oriented punishment but one was mistaken, participants thought that the target who actually received the action-oriented punishment would behave better in the future. When both targets believed they had received the person-oriented punishment but one was mistaken, participants also believed that the target who actually received the action-oriented punishment would behave better in the future (unlike in Study 2, where this mean was statistically indistinguishable from zero; -0.01). However, participants in the first of these

conditions were significantly more confident in this judgment than participants in the second. The same pattern appeared for learning from punishment judgments. That is, whatever produces the difference between the two one-mistake conditions likely does so by affecting judgments about learning, just as the communicative theory would predict.

These results suggest that participants' judgments about the degree to which the targets have learned from the punishments they received are based primarily on the targets' beliefs but may also incorporate other factors when beliefs do not provide a basis for drawing a distinction. In every condition in which at least one target believed he had received an action-oriented punishment, participants selected a target with such a belief as the one most likely to learn from punishment and behave better in the future. In the only condition where no target had such a belief, participants' judgments were indifferent between the two targets (in Study 2 and in the learning from punishment question in Study 3) or close to indifferent (in the future behavior question in Study 3). The only condition in which beliefs do not explain responses is the condition in which both targets believed they received an action-oriented punishment but one was mistaken. In that condition, participants thought the target who was not mistaken – who actually received the action-oriented punishment – would learn more from punishment and behave better in the future. This pattern suggests that participants place independent positive weight on two factors when making their learning and future behavior judgments: (1) whether a punished person believed he received an action-oriented punishment and (2) whether a punished person accurately believed he received an action-oriented punishment. Beliefs about the punishment received (the first factor) play the primary role in judgments about learning and future behavior, as the communicative theory predicts. Whether that belief is accurate only plays

the secondary role of differentiating between two punishment recipients who both believe they received an action-oriented punishment.

General Discussion

People understand punishment to serve a communicative function (Sarin et al., 2021). We investigated whether laypeople's perceptions of the message punishment sends differ based on whether the punishment is action-oriented or person-oriented. Across three experiments, participants expected people who received action-oriented punishments to behave better in the future than people who received person-oriented punishments. This expectation arose to the greatest degree when the punishment recipients were aware of the type of punishment they received. This expectation did not appear to depend on whether the state or the victim ultimately received the money paid out as punishment, suggesting that the results do not merely reflect participants' preferences for victim compensation. This expectation also did not appear to depend on the intentions of the punisher, suggesting that our participants did not interpret the imposition of a particular type of punishment as a signal of hidden information about the recipient's true self.

This pattern of results is most consistent with a communicative theory of punishment, coupled with a new distinction we have drawn in this paper: Some punishments are action-oriented and send a didactic message, while others are person-oriented and send a less didactic, and potentially ostracizing, message. Action-oriented punishment achieves its goals by undoing the harmful effects of the punishment recipient's norm violation, including by forcing them to make whole any victims. In doing so, action-oriented punishment both reflects community norms against harming others and forces the recipient to enact those norms. Our studies, like prior research, have found that people expect punishment that sends this type of didactic message to

give rise to more normative future behavior (Sarin et al., 2021). Person-oriented punishment, by contrast, achieves its goals by harming the punishment recipient in response to a norm violation. The structure of this punishment, which excludes the victim's participation, does not reflect why the recipient's action was wrong, suggest that harm can and should be mitigated, or require any type of acknowledgment from the recipient of the community's values. In this way, person-oriented punishment does not send a message of ongoing inclusion in the community or signal an interest in teaching the recipient how to adhere to the community's values. Instead, prior research suggests that people may view person-oriented punishment as appropriate for, and thus potentially implying to the recipient the existence of, a bad true self (Maffly-Kipp et al., 2022). Such a message may appear less likely to lead to normative future behavior.

The finding in Study 1 that people expect better future behavior from recipients of action-oriented punishment than from recipients of person-oriented punishment is susceptible to multiple explanations. People could simply prefer action-oriented punishment because it compensates the victim and say it leads to better consequences as a way of expressing approval. People could expect the state to inflict person-oriented punishments on individuals perceived to have a bad true self and treat the state's judgments as evidence for the recipient's actual true self – that is, assume the state's true self judgments are based on private, accurate information about the recipient's true capacity for improving their behavior going forward. However, neither of these explanations can account for our finding that people's expectations of future behavior depend mostly on what type of punishment the recipient believed he received, rather than the punishment he actually received or the punishment the state intended to impose, as shown in Studies 2 and 3.

This evidence that participants perceive a difference in the communicative content of action-oriented and person-oriented punishments sheds additional light on the way people use and interpret punishment. First, this difference helps explain under what conditions people think that a punishment has communicated a message that will be effective at changing future behavior. A harm inflicted for a potentially selfish motive is ambiguous as to whether the harm is a punishment at all, and a person-oriented punishment is ambiguous as to what the punisher is condemning and why (Sarin et al., 2021). Second, the judgments reflected in the present studies may help explain why adults and children often prefer restoration over punishment (Bicchieri & Maras, 2022; Maffly-Kipp et al., 2022; Riedl et al., 2015). Participants in these earlier studies may have imposed action-oriented rather than person-oriented punishments because they believed that action-oriented punishments would more effectively improve the future behavior of the person who violated the norm. A similar intuition may play some role in the related findings that children sometimes prefer compensation over punishment (Lee & Warneken, 2020; though see McAuliffe & Dunham, 2021, for a different view) or adults' preference for compensation when placed in the role of victim (FeldmanHall et al., 2014).

These findings may appear to be in some tension with the evidence that, in an economic game, recipients of person-oriented punishment subsequently engage in more prosocial behavior than recipients of action-oriented punishment (de Vel-Palumbo et al., 2023a). There are two possible ways to reconcile these findings. First, as de Vel-Palumbo and colleagues (2023a) acknowledge, in their study, the opportunity for prosocial behavior occurred immediately after the punishment. Thus, their finding was that participants who had just received action-oriented punishment – i.e., forced compensation of the other players in their game – provided less voluntary compensation to the same group of players immediately afterwards. In the present

studies, participants made predictions about the punishment recipients' future behavior in general and over a long period of time. These predictions could have been accurate without contradicting the findings of de Vel-Palumbo and colleagues (2023a). Second, the participants in the present study could have been mistaken about how punishment affects future behavior. That is, their predictions that action-oriented punishment would lead to better future behavior than person-oriented punishment in general and over long timeframe could be inaccurate. Further research would be required to determine which of these explanations is correct.

The current findings also bear on the philosophical debate surrounding the theory of pure restitution – that is, the argument that the state should only use restitution, a type of action-oriented punishment, to respond to criminal offenses[11] (Barnett, 1977; Boonin, 2008). One central point of contention in that debate is whether action-oriented punishment provides adequate deterrence. The present results show that laypeople think action-oriented punishments will lead to less crime and better future behavior than person-oriented punishments. Moreover, by providing additional evidence that people interpret purely monetary punishments to have communicative content (Sarin et al., 2021), these studies further demonstrate that laypeople disagree with the argument that only carceral penalties can communicate (e.g., Feinberg, 1965; Kahan, 1996).

Limitations and Future Directions

[11] There is ongoing philosophical debate about whether restitution counts as punishment (Boonin, 2008) and whether restitution encompasses orders to return secondary costs of a crime, such as the cost of inconvenience to the victim. These semantic debates may affect how one talks about the results of our studies but not their theoretical import, as we did not ask participants questions that assumed any particular definition of "punishment" or "restitution."

The present studies advance scientific understanding of the expressive function of punishment by providing evidence that people understand different types of punishment to send different messages to recipients. However, like all research, these studies have limitations.

First, in order to isolate the contrast between action-oriented and person-oriented punishments, the vignettes used in these studies all described punishments that were either purely action-oriented or purely person-oriented. However, real-world punishment can combine both types of punishment. (Although the evidence suggests that real judges impose arguably action-oriented alternative sanctions like community service less than laypeople would; Mott & Solomon, 2024.) Indeed, punishments for serious property crimes often include both incarceration (or another person-oriented punishment like probation) and mandatory restitution (e.g., United States Sentencing Commission, 2024, §§ 2B2.1, 5A, 5E1.1). The present studies do not address what inferences people draw from these types of combined sentences. However, the theoretical framework described in this paper suggests two possibilities. First, if people interpret action-oriented punishment to communicate a didactic message and person-oriented punishment to communicate an ostracizing message, then those two messages may conflict. In that case, the interpretation of the punishment will likely depend on which type of punishment is primary; for instance, laypeople would likely treat a sentence involving a long prison term and a small amount of restitution as a person-oriented punishment. Indeed, in the United States, people often discuss multi-component sentencing packages that include incarceration along with a variety of alternative punishments as if they included incarceration alone (Mott & Solomon, 2024). Second, if people interpret action-oriented punishment to communicate a didactic message and person-oriented punishment to communicate a merely ambiguous or unfocused message (Sarin et al., 2021), then the interpretation of the punishment will likely depend primarily on the amount of

action-oriented punishment. Future research should explore this question by eliciting judgments about learning from punishment and expectations of future behavior for people who received multi-component sentences.

Second, the vignettes used in these studies did not explicitly describe the judges communicating any messages to the people they were sentencing aside from the message inferred from the punishment itself. One may wonder whether explicit messages from the judges – taking either a didactic or an ostracizing form – would overwhelm any message contained in the punishment itself. This hypothesis is plausible, as an explicit message may be more powerful than an implicit message. However, there is some existing evidence in the other direction. In de Vel-Palumbo and colleagues' (2023a) Study 4, half of participants received a punishment we would describe as action-oriented and half received a punishment we would describe as person-oriented. When the punishment was delivered, participants in both conditions received the same reasons, which described their failure to donate all their points to the public good as "egoistic" because it made their team members "worse off" (de Vel-Palumbo et al., 2023b, Supplementary Material, p. 31). That is, participants in both conditions received a didactic explanation of the punishment. Nevertheless, participants drew different motive inferences in the two punishment conditions, accepted the punishment to different degrees, and behaved differently after the punishment. This finding suggests that the message sent by the punishment overwhelmed the content of the verbal message. However, future research could investigate this question by comparing learning and future behavior judgments between conditions with just an implicit message (i.e., punishment alone) and conditions with both an implicit and an explicit message (e.g., punishment paired with a lecture from a judge).

Third, the studies reported in this paper elicited judgments about people who had committed property crimes of varying severity. We used vignettes about property crimes because they allowed us to ensure that the victim was fully or near-fully compensated for the harm. Harms from physical violence, including potentially long-lasting psychological trauma, are necessarily more difficult to quantify and may prove impossible to repair (Boonin, 2008). Indeed, people's responses to a wide variety of hypothetical criminal violations indicate that punishing the perpetrator is comparatively more important than compensating the victim in cases of serious crime, particularly violent crime of types that cause permanent physical or psychological injury to the victim (Heffner & FeldmanHall, 2019). This finding suggests that the perceived advantage of action-oriented punishment for improving future behavior may fade as the severity of the crime increases. To investigate further how laypeople understand the connection between a punishment's communicative content and its effectiveness, future research should explore whether these results continue to hold for crimes that inflict physical harms.

Fourth, the vignettes used in these studies described recipients of criminal punishment who had the means to pay the fines and restitution imposed upon them, which makes them relatively affluent. Many people involved in the justice system are not in this position, which means that sentences requiring monetary payment can hang over them for years, impairing their ability to earn a living (Barkow, 2019; Western, 2018). In such cases, monetary penalties may be counterproductive to producing better future behavior. For this reason, many restorative justice proposals eschew restitution and instead rely on mechanisms of restoration that involve investments of time rather than financial resources (Brooks, 2012). Comparing these types of action-oriented punishments to person-oriented punishments is more challenging than comparing two different types of monetary payment, which differ only in their recipients. However, future

research should investigate how laypeople understand the communicative content and effectiveness of these types of action-oriented punishments.

The present findings also point toward an interesting question for future research on the interaction between punishment and learning – namely, whether participants are correct that action-oriented punishment leads to better future behavior than person-oriented punishment. Multiple lines of converging evidence suggest they may be.

First, research on instrumental conditioning suggests that action-oriented punishments are more effective at teaching recipients how to behave. Both action-oriented punishment and person-oriented punishment satisfy the definition of punishment used in this literature – each one involves either adding a "noxious stimulus" (Solomon, 1964, p. 239), removing a reinforcer (Byrne & Poling, 2017), or both, and they can therefore change future behavior by making the triggering behavior less frequent (for related conceptualizations of punishment, see Cushman et al., 2019; Jordan et al., 2016; Marshall & McAuliffe, 2022). Although punishment rapidly drops off in effectiveness in seconds or minutes among non-human animals (Byrne & Poling, 2017; Kamin, 1959), research in humans has shown that punishment can affect behavior even after substantial delays when paired with an explanation of the specific reasons for the punishment (Verne, 1977) or when clearly contingent on a target behavior (Marciano et al., 2015; for a review, see Meindl & Casey, 2012).

Action-oriented punishment may therefore improve future behavior more than person-oriented punishment by expressing the type of specific reasons necessary to change behavior after a time delay. Action-oriented punishment involves undoing the harm of the crime, so it often resembles the punished behavior. For example, someone who steals a bike takes someone else's bike with the intention of retaining it. An action-oriented punishment for that theft would

involve undoing that harm to the extent possible – e.g., a transfer between the same two people of identical property (the bike itself) or property of similar value (compensation for the bike), also with the intention that this transfer be permanent. This resemblance between crime and punishment helps resolve any ambiguity about what triggered the punishment (e.g., the punishment recipient's specific theft behavior) as well as *why* the crime was wrong. By undoing the harm previously inflicted, action-oriented punishment re-enacts the transfer of property involved in the crime and thereby provides reasons for the punishment grounded in the recipient's behavior. Person-oriented punishment, by contrast, merely harms the recipient, which is more likely to lead to ambiguity about the reasons for its imposition[12] – in particular, whether the reason for the punishment really is the person' norm-violating behavior as opposed to some aspect of the person's identity (e.g., having a bad true self).

Second, a substantial body of research on the criminal justice system provides evidence that people who receive action-oriented punishments are, in fact, less likely to re-offend than people who receive person-oriented punishments. Studies examining the effects of restorative justice approaches, which we have noted are similar to action-oriented punishment, show that they lead to significantly lower rates of re-offense than participation in the criminal justice system (Latimer, 2005; Shem-Tov et al., 2021). Conversely, rates of re-offense after incarceration, the most common form of person-oriented punishment, are high and careful econometric analysis reveals no or little causal effect of additional imprisonment on reducing those rates (Kuziemko, 2013; Rose & Shem-Tov, 2021). However, further laboratory and observational studies are needed to determine whether and when lay perceptions about the

[12] The difficulties involved in resolving this type of ambiguity are related to the credit assignment problem that arises in reinforcement learning (Niv, 2009; Wilson & Niv, 2012).

comparative effectiveness of action- and person-oriented punishments at bringing about normative future behavior are accurate.

Conclusion

The present studies suggest that laypeople draw inferences not just from the fact that a person experienced punishment but also from the *type* of punishment. They view action-oriented punishments as more effective at bringing about normative future behavior than person-oriented punishments. These views appear to arise, at least in part, from an understanding that punishment is expressive and that its communicative content depends on characteristics of the punishment. Action-oriented punishment, which in form and effect reflects both the nature of the recipient's norm violation and why that violation was wrong, connects the harm of punishment more closely to the recipient's target behavior and forces the recipient to enact the community's values. Person-oriented punishment, by contrast, expresses a less behavior-specific, and thus more ambiguous, message. In other words, people view punishments focused on undoing the harm to the victim as more effective at reducing crime than punishments that merely harm the recipient.

Open Practices

Vignettes, measures, materials, data, and analysis code for all studies appear at: https://osf.io/nxw24/?view_only=b92f21261c8e4804857bccd6099af96b. Pre-registrations describing exclusion criteria and statistical analyses appear at the following links:

Study 1: https://aspredicted.org/3J2_XKW

Study 2: https://aspredicted.org/MWD_S24

Study 3: https://aspredicted.org/2FC_V3L

References

Alexander, M. (2010). *The new Jim Crow*. New Press.

Anderson, J.R. (2000). *Learning and memory: An integrated approach* (2nd Ed.). John Wiley & Sons, Inc.

Aronfreed, J. (1968). *Conduct and conscience: The socialization of internalized control over behavior*. Academic Press.

Barkow, R.E. (2019). *Prisoners of politics: Breaking the cycle of mass incarceration*. Harvard University Press.

Barnett, R. (1977). Restitution: A new paradigm of criminal justice. *Ethics, 4*, 279-301, https://doi.org/10.1086/292043

Bicchieri, C., & Maras, M. (2022). Intentionality matters for third-party punishment but not compensation in trust games. *Journal of Economic Behavior & Organization, 197*, 205-220, https://doi.org/10.1016/j.jebo.2022.02.026

Boonin, D. (2008). *The problem of punishment*. Cambridge University Press.

Braithwaite, J. (2002). *Restorative justice & responsive regulation*. Oxford University Press.

Bregant, J., Shaw, A., & Kinzler, K. D. (2016). Intuitive jurisprudence: Early reasoning about the functions of punishment. *Journal of Empirical Legal Studies, 13*, 693-717.

Brooks, T. (2012). *Punishment*. Routledge.

Byrne, T. & Poling, A. (2017). Behavioral effects of delayed timeout from reinforcement. *Journal of Experimental Analysis of Behavior, 107*, 208-217.

Caruso, G.D. (2021). *Rejecting retributivism: Free will, punishment, and criminal justice*. Cambridge University Press.

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. Social Psychological and Personality Science, 11, 464-473, https://doi.org/10.1177/1948550619875149

Christensen, R. H. B. (2015). ordinal - Regression Models for Ordinal Data. R package version 2022.11-16. https://cran.r-project.org/package=ordinal.

Christy, A. G., Schlegel, R. J., & Cimpian, A. (2019). Why do people believe in a "true self"? The role of essentialist reasoning about personal identity and the self. *Journal of Personality and Social Psychology, 117*, 386–416. https://doi.org/10.1037/pspp0000254

Cushman, F. A., Sarin, A., & Ho, M. (2019). Punishment as communication. In J. Doris & M. Vargas (Eds.), *The Oxford Handbook of Moral Psychology*. Oxford University Press.

Denver, M., Pickett, J. T., & Bushway, S. D. (2017). The language of stigmatization and the mark of violence: Experimental evidence on the social construction and use of criminal record stigma. *Criminology, 55*(3), 664-690. https://doi.org/10.1111/1745-9125.12145

Duff, R. A. (2001). *Punishment, communication, and community*. Oxford University Press.

Dunlea, J. P., & Heiphetz, L. (2021). Moral psychology as a necessary bridge between social cognition and law. *Social Cognition, 39*, 183-199.

Ewing, A.C. (1943). Punishment as viewed by the philosopher. *The Canadian Bar Review, 21*, 102-122.

Feinberg, J. (1965). The expressive function of punishment. *The Monist, 49*, 397-423, https://doi.org/10.5840/monist196549326

FeldmanHall, O., Sokol-Hessner, P., Van Bavel, J. J., & Phelps, E. A. (2014). Fairness violations elicit greater punishment on behalf of another than for oneself. *Nature Communications, 5*, 5306, https://doi.org/10.1038/ncomms6306

Funk, F., McGeer, V., & Gollwitzer, M. (2014). Get the message: Punishment is satisfying if the transgressor responds to its communicative intent. *Personality and Social Psychology Bulletin, 40*, 986-997, https://doi.org/10.1177/0146167214533130

Gahringer, R.E. (1960). Punishment as language. *Ethics, 71*, 46-48, https://doi.org/10.1086/291313

Gibson, L. (2021, July-August). Restoring justice. *Harvard Magazine*, https://www.harvardmagazine.com/2021/07/features-restorative-justice

Hampton, J. (1984). The moral education theory of punishment. *Philosophy & Public Affairs, 13*, 208-238.

Hampton, J. (1992). Correcting harms versus righting wrongs. *UCLA Law Review, 39*, 1659-1702.

Hart, H.L.A. (1968). *Punishment and responsibility: Essays in the philosophy of law*. Oxford University Press.

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd ed.). Guilford publications.

Heffner, J., & FeldmanHall, O. (2019). Why we don't always punish: Preferences for non-punitive responses to moral violations. *Scientific Reports, 9*, 1-13, https://doi.org/10.1038/s41598-019-49680-2

Heiphetz, L., Strohminger, N., Gelman, S. A., & Young, L. L. (2018). Who am I? The role of moral beliefs in children's and adults' understanding of identity. *Journal of Experimental Social Psychology, 78*, 210-219.

Hyatt, J. M., Andersen, S. N., & Chanenson, S. L. (2022). Nordic design: Embracing inspiration for reforming criminal fines. *Federal Sentencing Reporter, 34*, 155-165.

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

Kaeble, D. & Beatty, L.G. (2016). Correctional populations in the United States, 2015. *Bureau of Justice Statistics*, https://bjs.ojp.gov/library/publications/correctional-populations-united-states-2015

Kahan, D.M. (1996) What do alternative sanctions mean? *University of Chicago Law Review, 63*, 591-653, https://doi.org/10.2307/1600237

Kamin, L.J. (1959). The delay-of-punishment gradient. *Journal of Comparative and Physiological Psychology, 52*, 434–437.

Kirgios, E.L., Chang, E.H., Levine, E.E., Milkman, K.L., & Kessler, J.B. (2020). Forgoing earned incentives to signal pure motives. *Proceedings of the National Academy of Sciences, 117*(29), 16891-16897. https://doi.org/10.1073/pnas.2000065117

Klein, N., & O'Brien, E. (2016). The tipping point of moral change: When do good and bad acts make good and bad actors? *Social Cognition, 34*(2), 149-166. https://doi.org/10.1521/soco.2016.34.2.149

Kuziemko, I. (2013). How should inmates be released from prison? An assessment of parole versus fixed-sentence regimes. *The Quarterly Journal of Economics, 128*, 371-424.

Latimer, J., Dowden, C., & Muise, D. (2005). The effectiveness of restorative justice practices: A meta-analysis. *The Prison Journal, 85*, 127-144.

Lee, Y. E., & Warneken, F. (2020). Children's evaluations of third-party responses to unfairness: Children prefer helping over punishment. *Cognition, 205*, 104374, https://doi.org/10.1016/j.cognition.2020.104374

Lee, Y. E., Dunlea, J. P., & Heiphetz, L. (2023). Why Do God and Humans Punish? Perceived Retributivist Punishment Motives Hinge on Views of the True Self. *Personality and Social Psychology Bulletin*, 01461672231160027.

Lieberman, D.A. (2011). *Learning and memory* (2nd Ed.). Cambridge University Press.

Ma, Correll, & Wittenbrink (2015). The Chicago Face Database: A Free Stimulus Set of Faces and Norming Data. *Behavior Research Methods, 47*, 1122-1135. https://doi.org/10.3758/s13428-014-0532-5.

Maffly-Kipp, J., Rivera, G. N., Schlegel, R. J., & Vess, M. (2022). The effect of true self-attributions on the endorsement of retributive and restorative justice. *Personality and Social Psychology Bulletin, 48*, 1284-1297, https://doi.org/10.1177/01461672211027473

Maguire, E. R., Lowrey, B. V., & Johnson, D. (2017). Evaluating the relative impact of positive and negative encounters with police: A randomized experiment. *Journal of Experimental Criminology, 13*, 367-391. https://doi.org/10.1007/s11292-016-9276-9

Marciano, H., & Norman, J. (2015). Overt vs. covert speed cameras in combination with delayed vs. immediate feedback to the offender. *Accident Analysis & Prevention, 79*, 231-240.

Marshall, J., & McAuliffe, K. (2022). Children as assessors and agents of third-party punishment. *Nature Reviews Psychology, 1*, 334-344.

McAuliffe, K., & Dunham, Y. (2021). Children favor punishment over restoration. *Developmental Science, 24*, e13093, https://doi.org/10.1111/desc.13093

Meindl, J. N., & Casey, L. B. (2012). Increasing the suppressive effect of delayed punishers: a review of basic and applied literature. *Behavioral Interventions, 27*, 129-150.

Mott, C., & Solomon, L. H. (2024). Alternative punishments: How laypeople and judges impose alternative noncarceral sanctions. Psychology, Public Policy, and Law, 30(3), 326–347. https://doi.org/10.1037/law0000420

Murray, S., Krasich, K., Irving, Z., Nadelhoffer, T., & De Brigard, F. (2023). Mental control and attributions of blame for negligent wrongdoing. *Journal of Experimental Psychology: General, 152*, 120-138.

Nahmias, E., & Aharoni, E. (2017). Communicative theories of punishment and the impact of apology. In C. Surprenant (Ed.) *Rethinking Punishment in the Era of Mass Incarceration* (pp. 144-161). Routledge.

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science, 39*, 96-125, https://doi.org/10.1111/cogs.12134

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology, 53*, 139-154, https://doi.org/10.1016/j.jmp.2008.12.005

People v. Davis, 958 P.2d 1083 (Cal., 1998)

Primoratz, I. (1989). Punishment as language. *Philosophy, 64*, 187-205, https://www.jstor.org/stable/3751407

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2015). Restorative justice in children. *Current Biology, 25*, 1731-1735, https://doi.org/10.1016/j.cub.2015.05.014

Rose, E. K., & Shem-Tov, Y. (2021). How does incarceration affect reoffending? estimating the dose-response function. *Journal of Political Economy, 129*, 3302-3356.

Sarin, A., Ho, M. K., Martin, J. W., & Cushman, F. A. (2021). Punishment is organized around principles of communicative inference. Cognition, 208, 104544, https://doi.org/10.1016/j.cognition.2020.104544

Sered, D. (2019). *Until we reckon: Violence, mass incarceration, and a road to repair*. The New Press.

Shem-Tov, Y., Raphael, S., & Skog, A. (2021). Can restorative justice conferencing reduce recidivism? Evidence from the make-it-right program (No. w29150). *National Bureau of Economic Research Working Paper*, https://www.nber.org/papers/w29150.

Solomon, R. L. (1964). Punishment. *American Psychologist, 19*, 239–253. https://doi.org/10.1037/h0042493

State v. Chapland, 901 A. 2d 351 (N.J. 2006)

Strohminger, N., Knobe, J., & Newman, G. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science, 12*, 551-560, https://doi.org/10.1177/1745691616689495

Tobia, K. P. (2016). Personal identity, direction of change, and neuroethics. *Neuroethics, 9*, 37–43. https://doi.org/10.1007/s12152-016-9248-9

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science, 10*(1), 72-81. https://doi.org/10.1177/1745691614556679

United States v. Starks, 861 F.3d 306 (1st Cir. 2017)

United States Sentencing Commission (2024). *2024 guidelines manual annotated*. https://www.ussc.gov/guidelines/2024-guidelines-manual-annotated

de Vel-Palumbo, M., Twardawski, M., & Gollwitzer, M. (2023a). Making sense of punishment: Transgressors' interpretation of punishment motives determines the effects of sanctions. *British Journal of Social Psychology, 62*(3), 1395-1417. https://doi.org/10.1111/bjso.12638

de Vel-Palumbo, M., Twardawski, M., & Gollwitzer, M. (2023b, March 1). Punishment and motive attributions. https://doi.org/10.17605/OSF.IO/3TAYF

Verna, G. B. (1977). The effects of four-hour delay of punishment under two conditions of verbal instruction. *Child Development, 48*, 621-624.

Western, B. (2018). *Homeward: Life in the year after prison*. Russell Sage Foundation.

Wickham, H., Miller, E., Smith, D., Posit Software (2023). haven - Import and export 'SPSS', 'Stata' and 'SAS' files. R package version 2.5.4. https://cran.r-project.org/web/packages/haven/index.html

Wilson, R. C., & Niv, Y. (2012). Inferring relevance in a changing world. *Frontiers in Human Neuroscience*, 5, 189. https://doi.org/10.3389/fnhum.2011.00189

Wilson, J.P. & Rule, N.O. (2015). Facial trustworthiness predicts extreme criminal-semtencing outcomes. *Psychological Science, 26*(8), 1325-1331. https://doi.org/10.1177/0956797615590992

Zehr, H. (1990). *Changing lenses: A new focus for crime and justice*. Herald Press.