Supplementary Materials

Pilot Study

Before conducting the first study reported in the manuscript, we conducted a pilot study described below. The pilot study used Punishment Type (action-oriented vs. person-oriented) as a between-subjects independent variable and elicited non-comparative judgments about future behavior from the participants.

Methods

Participants

Based on similar prior research (Sarin et al., 2021), we expected that the effect of punishment orientation would be moderate in size ($f = 0.19$). Using alpha = 0.05 and power = 0.8, we therefore anticipated that we would need a total sample size of 219 and recruited 254 participants to account for potential exclusions. In accordance with the pre-registration (https://aspredicted.org/yvw6-nkxj.pdf), we excluded twenty-two 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 232 participants (M_{age} = 39.42 years; SD_{age} = 11.16 years; 45% female, 54% 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 $f = 0.18$ 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, 10% Asian or Asian-American, 5% Black or African-American, 3% multiracial, 0.4% Native American or Pacific Islander, and 1% “option not listed.” We asked about ethnicity separately from race, and 8% of participants self-identified as Hispanic or Latina/o. The sample’s highest level of education was distributed as follows: 13% high school or

GED, 19% some college or university, 10% associate's degree, 45% bachelor's degree, 11% master's or professional degree, and 3% PhD.

Procedure

Each participant was randomly assigned to one of four cells in a 2 (Punishment Type: action-oriented vs. person-oriented) x 2 (Transgressor Race: Black vs. White) between-subjects design. Each participant read four vignettes about a man in his 20s of the assigned Transgressor Race who had committed one of four common crimes: Property destruction, burglary, larceny, or assault. Each participant received all four vignettes in counterbalanced order.

Participants received different information about what punishment the transgressor had received depending on the Punishment Type condition. In the person-oriented punishment condition, participants learned that the defendant had paid a fine equal to 25% more than the amount of the victim's loss (i.e., the value of the property stolen or the monetary cost of dealing with injuries, in the assault case), which would go into the court's general fund. In the action-oriented punishment condition, participants learned that the defendant had paid restitution to the victim equal to 25% more than the amount of the victim's loss. Each punishment involved depriving the defendant 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 comparable in age, race, and a range of physical properties, as well as on ratings of traits like how angry, threatening, or trustworthy the faces appeared. To 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.

Participants then responded to four counterbalanced items related the defendant's ability to change; these items differed somewhat from those used to measure expectations of future behavior in the main manuscript. The bracketed text in the items below was replaced with a short description of the transgression:

- Do you think this person can change whether or not he [transgresses], if he wants to?
- Do you think this person will always [transgress]?
- Many years from now, will this person still [transgress]?
- Do you think this person has [transgressed] before?

For each measure, participants made two judgments: a binary choice between the two defendants and a rating of confidence in their choice on a three-point Likert scale, from "Not at all sure" to "Very sure." We converted these two responses into a score on a scale from -2.5 to 2.5 by multiplying two variables coded as follows: For the binary choice, we coded a choice of the defendant who received the action-oriented punishment to behave better in the future as 1 and a choice of the other defendant (who received a person-oriented punishment) as -1; for the confidence ratings, the lowest confident was coded as 0.5 and the highest as 2.5, with one-unit increments between the three levels. Thus, each of the four items received a score where -2.5 reflected the lowest confidence that the defendant could change and +2.5 reflected the highest confidence that the defendant could change.

Following this main measure of interest, each participant responded to ten items adapted from Tyler and Fagan (2008) to elicit participants' views of the legitimacy of the court system. We planned to probe whether this measure moderated any significant effect of Punishment Type.

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.

Results

After confirming that the ability to change items had acceptable reliability within each between-subjects condition (αs > 0.85), we computed the dependent variable by averaging the scale scores across all items and vignettes. A 2 (Punishment Type: action-oriented vs. person-oriented) x 2 (Transgressor Race: Black vs. White) between-subjects ANOVA with this dependent variable showed a marginal main effect of Punishment Type, $F(1, 227) = 3.44$, $p = .065$, $f = 0.13$, no significant main effect of Transgressor Race, $F(1, 227) = 0.28$, $p = .597$, $f = 0.04$, and no significant interaction, $F(1, 227) = 1.96$, $p = .163$, $f = 0.09$.

To explore whether the assault vignette may have attenuated the effect because it differed from all the other property-related vignettes, we re-ran the analyses above without that vignette included. Reliability scores were lower without this vignette but still acceptable, αs > 0.83. The 2 (Punishment Type: action-oriented vs. person-oriented) x 2 (Transgressor Race: Black vs. White) between-subjects ANOVA with this dependent variable once again showed a marginal main effect of Punishment Type, $F(1, 227) = 3.24$, $p = .073$, $f = 0.12$, no significant main effect of Transgressor Race, $F(1, 227) = 0.01$, $p = .941$, $f < 0.001$, and no significant interaction, $F(1, 227) = 1.84$, $p = .105$, $f = 0.11$.

Discussion

Based on this pilot study, we concluded that any difference between the two punishment types in ability to change – which is closely related to predictions about future behavior – likely

had a small effect size. We used this effect size to power Study 2. However, because the effect appeared so small based on these results, we did not pre-register a directional hypothesis, as it seemed possible the comparative judgments would turn out differently.

This pilot study also led to changes in the materials and design of the later studies. We refined the vignettes and dependent variable items to make them uniform and increased the confidence scale from a three-point scale to a five-point scale to allow more variance in responses. Finally, we speculated that participants in a between-subjects design may not focus on the difference in punishment types, given all the other information in the image and vignette. Therefore, to focus attention on the distinction of interest, we converted the dependent variable to a comparative judgment between defendants who had received different types of punishment for the studies reported in the manuscript.

Study 3

Simulation for Power Analysis

This study used a 2 (Mistake for Person-Oriented Punishment Recipient: yes vs. no) x 2 (Mistake for Action-Oriented Punishment Recipient: yes vs. no) between-subjects design. The communicative theory predicts large main effects of similar size (and similar in size to the effect observed in Study 2) and, at most, a small interaction effect – if, e.g., the slightly higher than expected mean in the Study 2 mistake condition, which appeared to be due to some other process, resulted in something less than a complete inversion of the no-mistake condition mean for the two-mistakes condition. Thus, this study requires a sample large enough to detect the smallest plausible main effects, as well as an interaction effect of sufficiently large size that it would call the communicative account into question.

To determine what these effect sizes were, we created a simulated data set based on the results of Study 2. Given that Study 2 used three of the four conditions of this planned study, we drew the means and standard deviations for those conditions directly from the Study 2 results when drawing values for the simulation. To construct a plausible mean for the two-mistake condition, we subtracted the mean of the no-mistake condition from the mean responses across both one-mistake conditions. We assumed the same standard deviation in the two-mistake condition as in the no-mistake condition.

We then drew 1,000 simulations of this data and calculated the effect size of the main effects and the interaction effect each time. To ensure we powered Study 3 using a main effect size we could expect to see on the majority of occasions, we used the first quartile main effect size from this simulation (partial $\eta^2 \approx .14$). To ensure we powered Study 3 to detect any interaction effect larger than could arise by chance from the variation in the Study 2 responses, we powered the study to detect any interaction effect larger than the 95th percentile interaction effect in the simulation (partial $\eta^2 \approx .03$). Note that powering the study to detect such a small interaction effect means that it could, in fact, detect the first percentile main effect size in this simulation (partial $\eta^2 \approx .05$).