What people learn from punishment: joint inference of wrongness and punisher's motivations from observation of punitive choices

Setayesh Radkani (radkani@mit.edu)
Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139 USA

Rebecca Saxe (saxe@mit.edu)
Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139 USA

Abstract

Punishment is a cost imposed on a target, in response to an undesirable action. Yet choosing to punish also reveals information about the authority's own motives and values. We propose that observers jointly infer the wrongness of the action and the authority's motivations. Using hypothetical scenarios in unfamiliar societies, we experimentally manipulated observers' prior beliefs and measured human observers' inferences after observing punishment. These inferences were recapitulated in a formal model that inverts an intuitive causal model of authorities who make rational choices about punishment by weighing its costs and benefits (i.e. utilities). An essential component of this model, driving these inferences, is that legitimate authorities consider the utility of a proportional response to harmful actions, which depends on the balance between the wrongness of the act and the severity of the punishment.

Keywords: punishment; moral cognition; Bayesian inference; moral learning; theory of mind

Introduction

Parents, teachers and other authorities choose whether and how severely to punish undesirable actions. Observers of punishment, including the target, can learn about the actions from these choices: the more severe the punishment, the more wrong the action. Yet in light of every punishment, observers and targets also evaluate the motives and legitimacy of the authority who punished. An authority who punishes too harshly, or who fails to punish a serious violation, may lose legitimacy. The central hypothesis of the current work is that these two inferences are fundamentally interdependent. Observers simultaneously infer the wrongness of the action and the motives of the authority, by rational Bayesian joint inference.

Punishment is a cost imposed on a target, in response to an undesirable (i.e. wrong, harmful, norm-violating) action. The more harmful or serious the transgression, the more people judge punishment to be a more appropriate response than not punishing (Eriksson, Andersson, & Strimling, 2017). This core intuition is shared by people from more than 50 countries (Eriksson et al., 2021). People also choose harsher, more severe punishments for more wrong or harmful actions (Shao & Perlow, 2009; Buckholtz et al., 2015; Ginther et al., 2016; Heffner & FeldmanHall, 2019; Sznycer & Patrick, 2020). People's intuitive sense of the relative seriousness of different transgressions is correlated with the relative severity of punishments assigned to these transgressions in both modern and ancient legal codes (Sznycer & Patrick, 2020).

The strong correlation between the severity of punishment and the wrongness of the act makes punishment eminently informative for observers learning or uncertain about an unfamiliar action (Wiessner, 2020; Darley, 2009). After observing more severe punishments, people form stronger moral disapproval of previously unfamiliar actions (Mulder, 2018). For example, in both an economic game and a hypothetical scenario about cheating, a larger fine imposed led to stronger moral disapproval of the target act (Mulder, 2009). In the early 2000s, young adults who had never shared music files, and learned about sanctions for such copyright infringements, subsequently judged sharing music files to be more unethical the more severe the sanction (Depoorter & Vanneste, 2005).

Choosing to punish, like any social action, also reveals information about the authority's own motives and values (Radkani, Tenenbaum, & Saxe, 2022). In third-party punishment, the person who chooses whether to punish was not directly harmed by the initial action, and will not be directly personally benefited by the punishment. In these settings, punishment is typically more immediately costly than not punishing, whereas the benefits of punishment (e.g. deterrence) are diffuse (Panchanathan & Boyd, 2004). Consequently, many studies have documented the positive perceptions that observers have of people who are willing to punish: in both economic games and vignette studies, third-parties who choose to punish are judged as more trustworthy and less selfish (J. J. Jordan & Rand, 2020), more competent and and more moral (Gordon, Madden, & Lea, 2014; Gordon & Lea, 2016; Dhaliwal, Skarlicki, Hoegg, & Daniels, 2020; de Kwaadsteniet, Kiyonari, Molenmaker, & van Dijk, 2019; J. Jordan & Kteily, 2020; Tsai, Trinh, & Liu, 2022), and are more likely to be chosen as cooperation partners than people who choose not to punish the same transgression. Harsher punishments signal more trustworthiness (Batistoni, Barclay, & Raihani, 2022). By contrast, people who fail to punish transgressions are sometimes punished themselves (Martin, Jordan, Rand, & Cushman, 2019; Tsai, 2021). Ethnographic studies confirm that third parties gain status from punishing (Wiessner, 2020). Such positive reputation benefits of punishment may be key to the evolution of punishment (Panchanathan & Boyd, 2004; Santos, Rankin, & Wedekind, 2011; Raihani & Bshary, 2015; Okada, 2020).

People do not simply associate severe punishment with trustworthy, competent, moral individuals responding to seri-

ous harmful transgressions, though. In some contexts, third-parties who choose to punish are judged as *less* trustworthy and *more* selfish than those who choose not to punish (Strimling & Eriksson, 2014; Eriksson, Andersson, & Strimling, 2016; Heffner & FeldmanHall, 2019; Rai, 2022; Sun, Jin, Yue, & Ren, 2022). People who use more severe punishment receive more disapproval, particularly if they stand to gain directly from the punishment (Eriksson et al., 2016; Rai, 2022). Even without direct benefits, people who choose harsh punishments are sometimes inferred to be taking selfish advantage, motivated to gain resources or power (Raihani & Bshary, 2015, 2019; Redhead, Dhaliwal, & Cheng, 2021). Likewise, severe punishment does not always induce inferences that the act was a serious transgression. Severe punishment of academic cheating did not increase disapproval of cheating in students with low trust in the educational institution (Mulder, 2009). Young adults who had engaged in music file sharing saw file sharing as *less* unethical, the more severe the proposed sanctions (Depoorter & Vanneste, 2005).

We propose that all of these seemingly complex or contradictory results can be explained parsimoniously by modeling observers as making rational joint inferences of the wrongness of the act and the motives and legitimacy of the authority, inverting a Bayesian causal model of the authority's choice to punish (Baker, Saxe, & Tenenbaum, 2011). What observers learn from the same punishment can vary, depending on the value and relative certainty of their prior beliefs about wrongness and legitimacy. Perceiving an authority as legitimate includes judging that their power was acquired by an appropriate process, their decision making is impartial and unbiased, and their motives are benevolent and sincere (Tyler, 2006; Tyler, Goff, & MacCoun, 2015). Here we focus on and formalize one aspect of impartiality and benevolence: punishments should be proportional to the harm caused by the punished act.

The more observers have confident prior beliefs about a punished act, then from observed punishment, the more they should infer the authority's motives: an authority who punishes too harshly, or who fails to punish a serious violation, will be perceived as less legitimate and just (e.g. Tsai, 2021). By contrast, the more observers have confident prior beliefs that the authority is legitimate, then from observed punishment, the more they should infer the wrongness of the act. To test this proposal we (i) conducted a series of experiments in which we manipulated and measured participants' beliefs about wrongness and authority's legitimacy, before and after observing an authority's decision about punishment, and (ii) built and validated a computational model of these inferences as inverse planning.

Experiments

Participants

100 US-resident adults (51 female; age range 18-75 years) who indicated English as their first language were recruited on prolific.co.

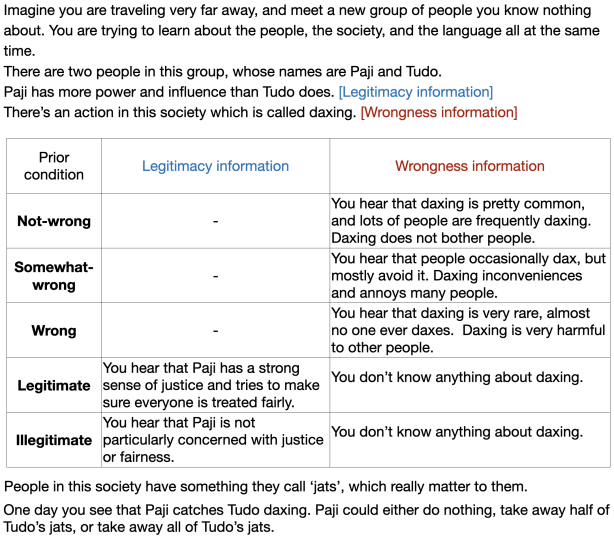

Figure 1: Example scenario

Materials

We developed five scenarios in hypothetical societies with unfamiliar people, actions, and punishments, to maximize experimental control of observers' priors. Each scenario introduces two people (an authority and a target), a novel target action, and finally three possible responses by the authority (do nothing, mildly punish, harshly punish). Depending on condition, we provide information about either the motivations of the authority (2 conditions, "Legitimate", "Illegitimate"), or the wrongness of the target act (3 conditions, "Not-wrong", "Somewhat-wrong", "Wrong"). An example scenario is shown in Figure 1. In a within-subject design, each participant read all five scenarios, one in each of the five conditions.

Manipulation check pilot

To test whether we successfully manipulated participants' priors, we conducted a pilot study. An independent set of raters (N=30) read the same scenarios and indicated their prior beliefs about wrongness and the authority's motivations. To measure both the value and the uncertainty of each individual's prior beliefs, we asked the participants to distribute 100 votes on the relevant scale for each question. These pilot data confirmed that participants had relatively flat priors for the wrongness of novel actions, prior to learning whether those actions were punished. These priors became sharply biased when the scenario described the action as common and harmless, rare and annoying, or very rare and seriously harmful. In the absence of information, participants were somewhat biased to believe that authorities are motivated by justice considerations, but priors on justice motivations were also substantially sharpened when the scenario contained explicit information about the authority (shaded densities in Figure 2). In sum, the pilot study confirmed that our scenarios manipu-

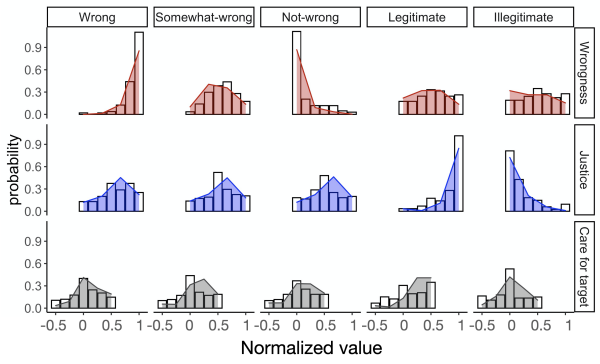

Figure 2: Distribution of prior beliefs in each experimental condition, before participants learned which response the authority chose. The shaded densities show the distribution of within-subject beliefs measured in the pilot experiment, and the white bars show the distribution of between-subjects prior beliefs in the main experiment.

lated both the value and the dispersion of participants' priors about novel actions and authorities.

Main experiment

In the main experiment, we used the same scenarios to evaluate how participants' beliefs change after learning about the authority's punitive choice. First, the participants indicate "without knowing what [the authority] chose to do" their prior beliefs about the wrongness of the target act ("In this society, how morally wrong is [target act]?", from 1: not at all to 7: extremely wrong), the authority's justice motivations ("In general, how much does [A] care about justice and acting fairly?", from 1: not at all to 7: extremely), and how much they care for the target ("In general, how does [A] feel when they see [T] get hurt?", from -3: very bad to 3: very good; reverse coded). In addition to the Likert scale, participants could also choose: "I don't know. All values are equally likely". Next, we repeated the same set of questions three times, if the authority chose to do each of the three possible responses (no punishment, mild punishment, harsh punishment), to measure the participants' posterior beliefs. The order of showing the three authority choices was randomized. Predictions, design and analysis plan were preregistered prior to data collection, at `tinyurl.com/47b26mc3`.

Analyses

For all analyses, we used mixed effects linear or logistic regression models, with the stated fixed effects and the full structure of random effects justified by the design (Barr, Levy, Scheepers, & Tily, 2013) unless the full model failed to converge. In cases of non-convergence, we pruned interaction random slopes, first from scenarios, then from participants. Then we removed the random effects altogether. To accommodate the effect of the variability in prior beliefs across the population, we used the update in beliefs after observing a specific punitive response as our main dependent variable. To do so, for each participant within each scenario, we subtracted the prior belief over each variable from their posterior judgements. "All values are equally likely" responses were replaced by the mid-point value of each scale, before carrying out the analyses. In some analyses, the average of responses for mild and harsh punishment is used to contrast "punish" vs "not-punish" responses.

Results

Punishment is typically intended to convey the undesirability and wrongness of the punished behaviour. Consistent with this function, participants in the current study inferred that punished acts were more wrong than not-punished actions (main effect of punish vs not-punish: β=0.758, std=0.080, df=99, p-value=1.76e-15). This effect was particularly clear when participants had no prior knowledge about the action, but believed that the punisher was a legitimate authority, motivated by justice. In the "Legitimate" condition, punished actions were inferred to be much more wrong than unpunished actions (main effect of punish vs not-punish: β=1.685, std=0.184, p-value<2e-16), and actions that received harsh punishment were inferred to be more wrong than actions that received relatively lenient punishment (main effect of harsh vs mild: β=0.930, std=0.175, p-value=3.05e-07).

On the other hand, punishment does not always convey the wrongness of the punished act to the same degree. The inference depends on both participants' prior beliefs about wrongness, and their prior beliefs about the authority's motivations. When participants have a strong prior belief that the target action was not wrong, they did not change this belief much, after observing punishment (interaction between response (punish vs not-punish) and condition ("Not-wrong" vs "Legitimate"): β=-1.220, std=0.252, p-value=2.16e-06; interaction between response (harsh vs mild) and condition, β=-0.700, std=0.236, p-value=0.0033). Even when participants had no prior knowledge about the action, if participants believed the authority was not motivated by justice, then punished actions were inferred to be only slightly more wrong than unpunished actions (interaction between response (punish vs not-punish) and condition ("Illegitimate" vs "Legitimate"): β=-0.885, std=0.236, p-value<0.001; interaction between response (harsh vs mild) and condition: β=-0.570, std=0.205, p-value=0.006). Indeed, participants reported that decisions of an illegitimate authority were not informative about the wrongness of the target action (main effect of "Illegitimate" vs "Legitimate" on probability of "I don't know" answers: β=1.994, std=0.277, p-value=6.50e-13).

We directly tested the prediction that observers' prior beliefs about the authority's motivations quantitatively determined the inference observers made from punitive responses. We used each participant's own reported prior in place of the experimenter label for each scenario. This analysis revealed a clear interaction between observer's priors about legitimacy and the punitive response in determining the wrongness inferences within "Legitimate" and "Illegitimate" conditions

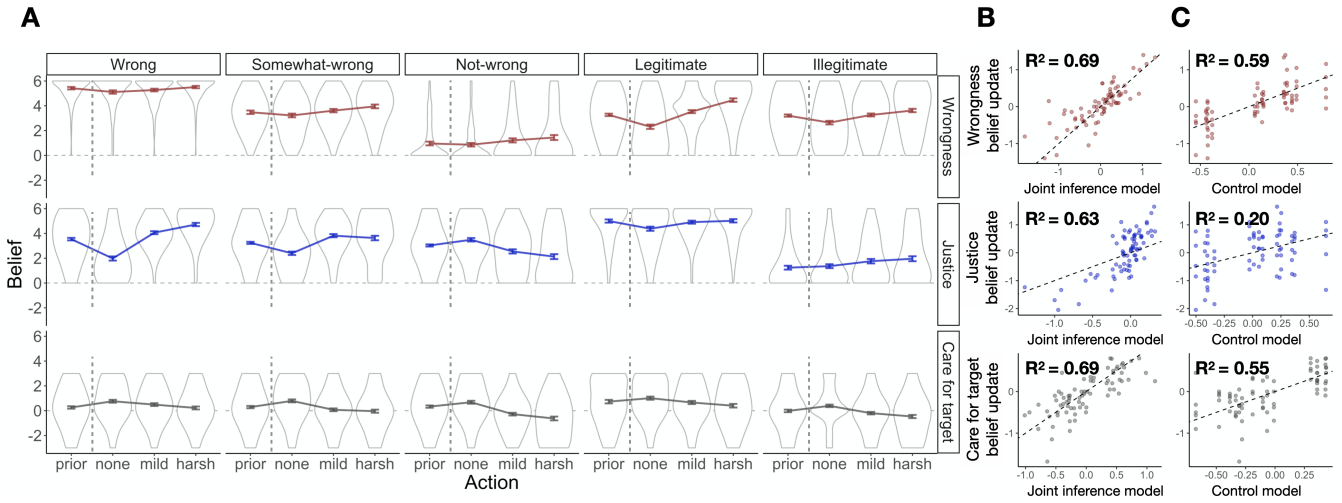

Figure 3: Human and model judgements. A) human prior and posterior beliefs across experimental conditions (error bars show the standard error of the mean); B) correlation of joint inference model predictions with human judgements on held-out scenarios. Each data point (75 in total) is the mean of the population in one prior-response condition for one scenario separately; C) correlation of the control model predictions with human judgements.

(β=0.174, std=0.023, df=439, p-value=4.15e-13). The higher participants' prior on authority's legitimacy, the more they updated their beliefs about action wrongness: negatively after observing no punishment, and positively after observing harsh punishment.

Yet, a decision whether and how to punish not only contains information about the punished action, but also about the authority's motives and values. In the current study, participants inferred that an authority who chose to punish was more motivated by justice than an authority who chose not to punish (main effect of punish vs not-punish: β=0.734, std=0.129, df=99, p-value=1.26e-07). This effect was particularly clear when participants had weak prior knowledge about the motives of the authority, and strong prior beliefs that the target action was harmful to many people. In the "Wrong" condition, authorities who punish were perceived to be much more motivated by justice than authorities who did not punish (β=2.416, std=0.292, df=4.03, p-value=0.0011). The biggest change in belief occurred when an authority did *not* punish a very harmful act. Mild and harsh punishment of harmful acts led to only small changes in inferred motives.

Again, though, punishment does not always imply justice motives in the authority. When participants had strong prior beliefs that the target action was harmless, then authorities who punished were perceived to be less motivated by justice than authorities who did not punish (interaction between response (punish vs not-punish) and condition ("Not-wrong" vs "Wrong"), β=-3.560, std=0.266, df=396, p-value<2e-16; main effect of punish vs not-punish in "Not-wrong" condition: β=-1.15, std=0.225, p-value=7.62e-07). Indeed, the inferences about authority's justice motivations depended on how the severity of punishment compared to the observers' beliefs about the wrongness of the target act (interaction between response (Mild vs Harsh) and condition — "Somewhat-wrong" vs "Wrong": β=-0.850, std=0.273, df=496, p-value=0.002, and "Not-wrong" vs "Wrong": β=-1.080, std=0.273, df=495, p-value=8.59e-05).

Observed punishment influenced inferences about how much the authority cares about the target, depending on prior beliefs about wrongness and justice motives. Severe punishment of not-wrong acts led observers to infer that the authority does not care about the target's suffering; but severe punishment of wrong acts did not support the same inference.

Overall, these results show that the same observed choice of whether, and how harshly, an authority punishes an action, can lead observers to make contrasting inferences. What observers learn from the same punitive response depends, with exquisite sensitivity, on the value and relative uncertainty of their prior beliefs about wrongness of the target act, as well as the authority's motivations.

Models

We propose that this pattern of inferences can be synthesized and captured, as joint inference inverting a causal model of authorities who make rational choices about punishment by weighing its costs and benefits (i.e. utilities).

Joint inference of wrongness and authority's motivations

Model specification In our model, authorities consider three utilities, when deciding how to react to a target action. First, the authorities consider the direct consequences of their response for the target (U_{target}), that is the punitive harm imposed on the target. Second, there is the utility of a proportional response to a harmful action in order to restore justice ($U_{justice}$). This utility depends on the balance between the

wrongness of the target act and the severity of the chosen punishment: $U_{justice}$ is highest when the punitive harm imposed on the target just offsets the harm caused by the target action, and is lower for punishments that are too lenient or too harsh. Finally, the authorities consider the direct costs and benefits of the chosen action for themself (U_{self}). Many models of punishment assume that punishment is directly costly or risky to the person (often, an equal status peer) who is punishing (J. J. Jordan, Hoffman, Bloom, & Rand, 2016). Here, the scenarios described decisions by authorities in which we expected punishment to carry little personal cost or risk. This design choice allows us to ignore U_{self} and focus on the dynamics of beliefs about wrongness and legitimacy.

Authorities weigh the positive utility of restoring justice against the negative utility of harming the target, to decide whether and how harsh to punish. The response to a specific target act varies depending on the authority's justice motivations and how much they care about the target. For instance, an authority with a strong weight on the utility of justice ($\alpha_{justice}$) would choose mild punishment for somewhat harmful actions and harsh punishment for very harmful actions, whereas an authority's response who does not particularly care about justice is less sensitive to the wrongness of the target act. An authority with a strong positive weight (α_{target}) on the target's utility would be more likely to choose mild or no punishment, while an authority with a strong negative α_{target} would choose harsh punishment, independent of the wrongness of the action. Therefore, the punitive response of the authority contains information about the authority's motivations and the wrongness of the target act. Indeed, by inverting this generative model, we capture how observers update their beliefs about both the wrongness of the action and the motives of the authority, based on the decision of whether and how harshly to punish.

To formalize these ideas, we write the authority's expected utility over each response 'p' as:

$$U_{total}(p) = \alpha_0 \alpha_{justice} U_{justice}(p) + \alpha_{target} U_{target}(p) \quad (1)$$

where U_{target} is more negative for harsher punishments and

$$U_{justice}(p) = \begin{cases} harshness(p) & harshness(p) \leq 0 \\ -\gamma\, harshness(p) & 0 < harshness(p) \end{cases} \quad (2)$$

$$harshness(p) = \eta_w\, wrongness - \eta_t\, U_{target}$$

α_0, γ, η_w and η_t are constants. Punitive responses are selected using a softmax decision rule:

$$P(p|wrongness, \boldsymbol{\alpha}, \boldsymbol{U}_{target}) \propto exp(\beta U_{total}(p)) \quad (3)$$

An observer holds a prior belief about the wrongness of the act, and how much the authority weighs each utility term ($\boldsymbol{\alpha}$). After observing a specific punitive response, the observer can then update their beliefs using these priors (i.e., $P(wrongness, \boldsymbol{\alpha})$) and their appraisals of how harmful each response is to the target (i.e., $\boldsymbol{U}_{target}$).

$$P(wrongness, \boldsymbol{\alpha}|p, \boldsymbol{U}_{target}) \propto P(p|wrongness, \boldsymbol{\alpha}, \boldsymbol{U}_{target}) P(wrongness, \boldsymbol{\alpha}) \quad (4)$$

where $P(p|wrongness, \boldsymbol{\alpha}, \boldsymbol{U}_{target})$ is the punisher's policy derived from equation (3).

Next, we test how this computational framework can capture the pattern of participants' inferences in our experiment.

Model simulation To test our model, we simulated it to explain the average population inferences within each scenario, each prior condition, and for each punitive response, separately (i.e., 5 scenarios $\times$ 5 conditions $\times$ 3 responses = 75 data points). We used a train-test procedure where we found the best fitting parameters (α_0, β and γ, minimizing mean squared error for wrongness, justice motives, and care for target) on 4 scenarios (i.e., 60 data points) and assessed model performance on the held-out scenario. This procedure was repeated for each of the 5 scenarios as test data. We report the average of model performance on the held-out data as the final performance measure.

We found the distribution of prior beliefs for each variable (wrongness, $\alpha_{justice}$, α_{target}), by fitting a beta distribution to the pool of participants' prior belief responses (normalized to [0,1]), within each scenario and prior condition separately. We considered "All values are equally likely" responses as a uniform distribution over the whole scale. The distribution of prior beliefs in the main study resembled the averaged prior distributions measured for every individual in the manipulation check pilot experiment (see Figure 2).

Another input to the model is how harmful the observers believe each response is to the target (i.e., $\boldsymbol{U}_{target}$). In our experiment, we measured U_{target} for each response separately, by asking "How much will [T] be hurt because of [A] doing [response]?", from 1: not at all to 7: very much. For each of the three responses in each scenario (using a novel means of punishment), we found the average U_{target} (normalized to [0,1]) across all conditions.

The last step in building the model was to determine how wrongness and U_{target} tradeoff to form beliefs about harshness (i.e., η_w and η_t), which according to equation 2 determines $U_{justice}$. We measured beliefs about harshness of each response by asking "[A] doing [response] was ... (-3: too lenient, 0: proportional/fair, 3: too harsh)". A regression model was fit to harshness judgements, with fixed effects of wrongness and U_{target}, using the data from conditions where wrongness beliefs were more stable ("Wrong", "Somewhat-wrong", "Not-wrong"). We assume people have a general concept of justice that does not vary by scenario; therefore, we estimated $\boldsymbol{\eta}$ in the training set and used the same parameters for the held-out scenario as well.

The joint inference model is able to explain a significant portion of variance in posterior judgements of all three variables simultaneously (average R^2 on held-out scenario = 0.95,

0.81, 0.70 for inferences of wrongness, justice motives and bias for or against the target, respectively). However, the nuances in belief updates could be masked by the larger shift in posterior beliefs as a result of manipulating the priors across different prior conditions; therefore, capturing the belief updates provides a stricter test of the model performance. As shown in Figure 3B, the model is able to explain a large portion of variance in belief updates as well.

Control models

To test whether the full structure of our model was necessary to fit human judgements, we asked how well we can capture people's beliefs using a model that makes inferences directly from the observed punitive harm imposed on the target, ignoring the prior beliefs and the trade-off between the authority's punitive motives. For this, we fit three linear regression models that were separately optimized to predict belief updates about wrongness, justice motives and care for target, using U_{target} (of each response within each scenario) as the regressor. To obtain model performance, we used a similar train-test procedure, where we fit the model on 4 scenarios and test on the held-out one, repeated for every scenario as test data.

As shown in Figure 3C, these models are able to capture the main effect of punishment severity on the inferences. However, although these models were optimized for each variable separately, they are less able to capture the nuances in belief updates across different prior conditions, compared to the joint inference model.

Discussion

Both in real life (Dölling, Entorf, Hermann, & Rupp, 2009; Nagin, 2013; Dhaliwal et al., 2020; Tsai, 2021) and laboratory settings (Mulder, 2009; Verboon & van Dijke, 2011; Eriksson et al., 2017), the same punishment can lead to contrasting and even contradictory consequences in terms of changing others' beliefs about undesirability of the act, as well as the motivations and legitimacy of the authorities. Prior research has begun to shed light on these discrepant findings by characterizing how beliefs about authorities' motivations affect how people change their beliefs about wrongness of the act (Tyler, 2006; Mulder, 2009; Verboon & van Dijke, 2011), and vice versa (J. Jordan & Kteily, 2020; Sarin, Ho, Martin, & Cushman, 2021). These studies typically focus on one component of these inferences at a time, and do not consider the interplay between them. Here we developed an experimental paradigm to control and study these inferences simultaneously, and showed that these two inferences indeed depend, with exquisite sensitivity, on one another. Both the value and uncertainty of beliefs about wrongness and authority's motivations affect what people learn from the same act of punishment and how they update their beliefs. Further, we proposed and validated a computational framework to explain such contrasting inferences parsimoniously, modeling observers as making rational joint inferences of wrongness and punisher's motivation by inverting a Bayesian causal model of how authorities make punitive decisions.

The joint inference framework is both general and flexible. Here, we instantiated a model where the authorities balance restoring justice against the negative consequences for the target. We manipulated observers' prior beliefs about the action, and the authority's justice motives. However, we did not manipulate observers' priors about the relationship between the authority and the target of punishment. When the authority and the target are allies, in-group members, or kin, then observers should have a prior that the authority values the target (high α_{target}) and is motivated to avoid harming them (Petersen, Sell, Tooby, & Cosmides, 2012; Molho, Tybur, Van Lange, & Balliet, 2020). In that case, harsh punishment should provide stronger evidence for the wrongness of the action, or for the legitimacy of the authority (in-group punishment communicates an impartial authority, Tsai, 2022). These inferences could be captured in the current model. We also did not manipulate the direct costs and benefits to the authority. Many prior experiments have found that observers make different inferences about the authority's motivations when the punishment (e.g. removing resources from the target) could directly benefit them (e.g. if the punisher receives part of the 'spoils'; Eriksson et al., 2016; Rai, 2022). Adding a utility for direct selfish costs and benefits to the current choice model could capture this pattern of inferences as well (Radkani et al., 2022).

Future research, using similar paradigms, could investigate whether authorities anticipate, and deliberately induce, the patterns of inferences that we documented here. For example, we found that observers inferred that an authority was more legitimate and motivated by justice, after observing the authority punish a wrong act. Authorities could therefore choose (public) punishment, communicatively, in order to boost their own perceived legitimacy (Tsai, 2021). A model of this choice could embed the current model of observer inferences recursively inside a model of authority's planning (Ho, Saxe, & Cushman, 2022; Radkani et al., 2022).

Reasoning in abstract hypothetical contexts can give insight into the cognitive processes evoked by real situations. Here we experimentally manipulated the observers' priors. Naturally occurring mismatches between the authority's, target's, or observers' priors could lead to rational but contrasting interpretations of the same punishment. We speculate that authorities often have strong prior beliefs in their own legitimacy and moral motivation. By contrast, observers and especially targets of punishment might be more confident about the wrongness (or not) of the target actions. For example, an authority might choose harsh punishment to express that the target action was very harmful; but the target might infer that the authority is biased against her. Conversely, an authority might choose lenient punishment to express that the target action occurred in extenuating circumstances; but observers might infer that the authority is not sufficiently motivated by justice. Thus, characterising how people jointly infer wrongness and legitimacy can illuminate why real world punishment attempts may fail or even backfire.

Acknowledgments

This work was supported by the Patrick J. McGovern Foundation grant.

References

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33).

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255–278.

Batistoni, T., Barclay, P., & Raihani, N. J. (2022). Third-party punishers do not compete to be chosen as partners in an experimental game. *Proceedings of the Royal Society B*, *289*(1966), 20211773.

Buckholtz, J. W., Martin, J. W., Treadway, M. T., Jan, K., Zald, D. H., Jones, O., & Marois, R. (2015). From blame to punishment: disrupting prefrontal cortex activity reveals norm enforcement mechanisms. *Neuron*, *87*(6), 1369–1380.

Darley, J. M. (2009). Morality in the law: The psychological foundations of citizens' desires to punish transgressions. *Annual Review of Law and Social Science*, *5*, 1–23.

de Kwaadsteniet, E. W., Kiyonari, T., Molenmaker, W. E., & van Dijk, E. (2019). Do people prefer leaders who enforce norms? reputational effects of reward and punishment decisions in noisy social dilemmas. *Journal of Experimental Social Psychology*, *84*, 103800.

Depoorter, B., & Vanneste, S. (2005). Norms and enforcement: The case against copyright litigation. *Or. L. Rev.*, *84*, 1127.

Dhaliwal, N. A., Skarlicki, D. P., Hoegg, J., & Daniels, M. A. (2020). Consequentialist motives for punishment signal trustworthiness. *Journal of Business Ethics*, 1–16.

Dölling, D., Entorf, H., Hermann, D., & Rupp, T. (2009). Is deterrence effective? results of a meta-analysis of punishment. *European Journal on Criminal Policy and Research*, *15*, 201–224.

Eriksson, K., Andersson, P. A., & Strimling, P. (2016). Moderators of the disapproval of peer punishment. *Group Processes & Intergroup Relations*, *19*(2), 152–168.

Eriksson, K., Andersson, P. A., & Strimling, P. (2017). When is it appropriate to reprimand a norm violation? the roles of anger, behavioral consequences, violation severity, and social distance. *Judgment and decision making*, *12*(4), 396–407.

Eriksson, K., Strimling, P., Gelfand, M., Wu, J., Abernathy, J., Akotia, C. S., ... others (2021). Perceptions of the appropriate response to norm violation in 57 societies. *Nature communications*, *12*(1), 1481.

Ginther, M. R., Bonnie, R. J., Hoffman, M. B., Shen, F. X., Simons, K. W., Jones, O. D., & Marois, R. (2016). Parsing the behavioral and brain mechanisms of third-party punishment. *Journal of Neuroscience*, *36*(36), 9420–9434.

Gordon, D. S., & Lea, S. E. (2016). Who punishes? the status of the punishers affects the perceived success of, and indirect benefits from,"moralistic" punishment. *Evolutionary Psychology*, *14*(3), 1474704916658042.

Gordon, D. S., Madden, J. R., & Lea, S. E. (2014). Both loved and feared: Third party punishers are viewed as formidable and likeable, but these reputational benefits may only be open to dominant individuals. *PLoS One*, *9*(10), e110045.

Heffner, J., & FeldmanHall, O. (2019). Why we don't always punish: Preferences for non-punitive responses to moral violations. *Scientific reports*, *9*(1), 1–13.

Ho, M. K., Saxe, R., & Cushman, F. (2022). Planning with theory of mind. *Trends in Cognitive Sciences*.

Jordan, J., & Kteily, N. (2020, Mar). *Reputation fuels moralistic punishment that people judge to be questionably merited.* PsyArXiv. Retrieved from psyarxiv.com/97nhj doi: 10.31234/osf.io/97nhj

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

Jordan, J. J., & Rand, D. G. (2020). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of personality and social psychology*, *118*(1), 57.

Martin, J. W., Jordan, J. J., Rand, D. G., & Cushman, F. (2019). When do we punish people who don't? *Cognition*, *193*, 104040.

Molho, C., Tybur, J. M., Van Lange, P. A., & Balliet, D. (2020). Direct and indirect punishment of norm violations in daily life. *Nature communications*, *11*(1), 3432.

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

Mulder, L. B. (2018). When sanctions convey moral norms. *European Journal of Law and Economics*, *46*, 331–342.

Nagin, D. S. (2013). Deterrence: A review of the evidence by a criminologist for economists. *Annu. Rev. Econ.*, *5*(1), 83–105.

Okada, I. (2020). A review of theoretical studies on indirect reciprocity. *Games*, *11*(3), 27.

Panchanathan, K., & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, *432*(7016), 499–502.

Petersen, M. B., Sell, A., Tooby, J., & Cosmides, L. (2012). To punish or repair? evolutionary psychology and lay intuitions about modern criminal justice. *Evolution and Human Behavior*, *33*(6), 682–695.

Radkani, S., Tenenbaum, J., & Saxe, R. (2022). Modeling punishment as a rational communicative social action. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Rai, T. S. (2022). Material benefits crowd out moralistic

punishment. *Psychological Science*, *33*(5), 789–797.
Raihani, N. J., & Bshary, R. (2015). The reputation of punishers. *Trends in ecology & evolution*, *30*(2), 98–103.
Raihani, N. J., & Bshary, R. (2019). Punishment: one tool, many uses. *Evolutionary Human Sciences*, *1*, e12.
Redhead, D., Dhaliwal, N., & Cheng, J. T. (2021). Taking charge and stepping in: Individuals who punish are rewarded with prestige and dominance. *Social and Personality Psychology Compass*, *15*(2), e12581.
Santos, M. d., Rankin, D. J., & Wedekind, C. (2011). The evolution of punishment through reputation. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1704), 371–377.
Sarin, A., Ho, M. K., Martin, J. W., & Cushman, F. A. (2021). Punishment is organized around principles of communicative inference. *Cognition*, *208*, 104544.
Shao, R., & Perlow, R. (2009). Effects of perceived responsibility, injury severity, and injury target on discipline severity. *Human performance*, *23*(1), 41–57.
Strimling, P., & Eriksson, K. (2014). Regulating the regulation: Norms about punishment. *Reward and punishment in social dilemmas*, 52–69.
Sun, B., Jin, L., Yue, G., & Ren, Z. (2022). Is a punisher always trustworthy? in-group punishment reduces trust. *Current Psychology*, 1–11.
Sznycer, D., & Patrick, C. (2020). The origins of criminal law. *Nature human behaviour*, *4*(5), 506–516.
Tsai, L. L. (2021). *When people want punishment: Retributive justice and the puzzle of authoritarian popularity*. Cambridge University Press.
Tsai, L. L., Trinh, M., & Liu, S. (2022). What makes anticorruption punishment popular? individual-level evidence from china. *The Journal of Politics*, *84*(1), 602–606.
Tyler, T. R. (2006). Psychological perspectives on legitimacy and legitimation. *Annu. Rev. Psychol.*, *57*, 375–400.
Tyler, T. R., Goff, P. A., & MacCoun, R. J. (2015). The impact of psychological science on policing in the united states: Procedural justice, legitimacy, and effective law enforcement. *Psychological science in the public interest*, *16*(3), 75–109.
Verboon, P., & van Dijke, M. (2011). When do severe sanctions enhance compliance? the role of procedural fairness. *Journal of Economic Psychology*, *32*(1), 120–130.
Wiessner, P. (2020). The role of third parties in norm enforcement in customary courts among the enga of papua new guinea. *Proceedings of the National Academy of Sciences*, *117*(51), 32320–32328.