Modeling punishment as a rational communicative social action

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

Joshua B. Tenenbaum (jbt@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

When deciding whether and how to punish, people consider not only the potential direct consequences, but also, how their choice will affect observers' judgements about the values and motives underlying the choice. We formalize the decision to punish as a rational communicative social action (RCSA). The model generates rational decisions to punish, incorporating anticipated observers' judgements obtained from a recursive model of inference using an intuitive theory of mind. Using this model, we synthesize patterns of human punishment from recently published papers. RCSA thus offers a formal model of the cognitive process that humans use to balance preferences for how they are perceived, with other goals for punishing.

Keywords: punishment; reputation; pragmatics; social cognition; moral cognition; communication; Bayesian model; planning; theory of mind

Introduction

Imagine you are a student living in a dorm, and you happen to see some of the other students in your dorm cheating on a take-home exam. Would you report them, so they fail the exam? Your decision whether or not to punish the cheaters likely depends on many features of the situation, including whether the cheating was deliberate, whether it caused harm, and whether punishing the cheaters is likely to teach them, or others, not to cheat in future. Here we focus on another key input: what will other people infer about you, from your decision? Whereas traditional game-theoretic models of punishment focus on the direct costs and benefits of punishments, recent experiments suggest that human punitive decisions are sensitive to how the punisher desires to be perceived. Our overall goal is to characterise the cognitive processes that happen, in human minds, while making punitive decisions.

Evolutionary models show that costly third-party punishment can be adaptive, if observers preferentially cooperate with or trust punishers, in subsequent interactions (Panchanathan & Boyd, 2004; Santos, Rankin, & Wedekind, 2011; Raihani & Bshary, 2015; Okada, 2020). Yet these models make no commitment about how an agent actually chooses to punish. Here we offer a model of the cognitive process underlying the decision to punish in humans. We propose that people rationally choose whether and how to punish, using a recursive model of observers' inferences. We develop a model framework that integrates the Bayesian Theory of Mind (BToM) model of inverse planning (C. L. Baker, 2011), with the rational speech act (RSA) model of pragmatic communication (Frank & Goodman, 2012; Goodman & Frank, 2016), and thus model decisions to punish as *rational communicative social actions* (RCSA, Figure 1).

The central premise of RCSA is that, in addition to the direct consequences of their actions, people value how their values and motives are perceived by observers of those actions. When choosing an action, people recursively model the inference that observers would make from each possible choice. Actions that generate the desired inference in observers are more valuable. For example, in some situations, people may want to be seen as unselfish, and so may avoid actions that they expect observers to perceive as selfishly motivated.

RCSA provides a principled quantitative framework for incorporating the value of anticipated observer inferences into a model of socially meaningful actions. Modelling the decision of whether and how to punish as a communicative social action allows us to capture, in one framework, five recently reported patterns in human punishment. Beyond capturing these existing findings, the RCSA framework predicts the conditions that evoke, or modulate, these patterns of behaviour. In the discussion, we consider the broader range of phenomena that could be captured by a more general model of punishment as RCSA.

Relationship to prior work

Punishment can be used to communicate in distinct ways. Prior research has characterized how punishment can be used to communicate *about the transgression* that evoked the punishment (Bregant, Shaw, & Kinzler, 2016). Directly communicating which actions are desirable, or undesirable, using rewards and punishments, is more efficient if both the punisher and the target assume that these signals are generated pedagogically (Ho, Cushman, Littman, & Austerweil, 2019; Sarin, Ho, Martin, & Cushman, 2021). Indeed, a long tradition of philosophical and legal theory argues that punishment is not intended just as direct negative reinforcement of an action, but rather as an expression of a community's disapproval and rejection of that action (Feinberg, 1965; Primoratz, 1989; Sunstein, 1996; Mulder, 2018).

Here we are concerned with a separate question: how punishment can be used to communicate about *the values and motives of the punisher*. Similar models have been used in previous work to capture how the desire to appear impartial can influence a person's choice to distribute unequal financial payouts (Kleiman-Weiner, Shaw, & Tenenbaum, 2017);

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

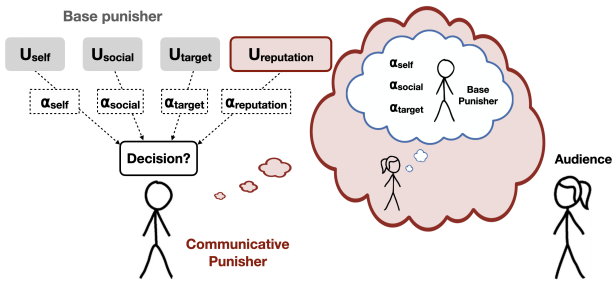

Figure 1: RCSA framework applied to punishment. The base punisher weighs ($\boldsymbol{\alpha}$) the consequences (U) for herself, the social good, and the target. The communicative punisher weighs, in addition, whether observers will make the desired inference about the punisher's internal values.

how the desire to appear considerate of other's feelings can influence the choice of negative expressions in polite speech (Yoon, Tessler, Goodman, & Frank, 2020); or how people balance the value of learning new skills against the desire to appear competent at familiar skills (Yoon, MacDonald, Asaba, Gweon, & Frank, 2018; Asaba & Gweon, 2019).

RCSA framework

First, we model the choice to punish without communication ("base punisher"). Then, we add recursive reasoning to model the choice to punish, given communicative goals ("communicative punisher").

Base punishment, without communication

To begin, we define punishment as an action with three types of direct consequences. First, a punishment imposes a cost on the target of punishment. Examples of punishment range from punishments imposed by the criminal justice system (fines, jail time), to those imposed by parents (time out, no dessert), to those imposed by anonymous strangers in lab experiments (endowment reduced by 5 points), but all share the central feature of an imposed cost.

Second, a punishment is expected to achieve a future social benefit, typically described in terms of specific deterrence (improving the future behaviour of the target) or general deterrence (improving the future behaviour of observers). How, and how effectively, punishments actually achieve their intended benefits is disputed (Sunstein, Kahneman, & Schkade, 1997; Dölling, Entorf, Hermann, & Rupp, 2009; Nagin, 2013). Here we consider situations in which punishers believe, and expect others to believe, that punishing will achieve some social benefit.

Third, choosing to punish has direct consequences for the punisher. Here we focus on anonymous, third-party punishment, in which the potential punisher was not negatively affected by the target's initial action and does not expect to directly interact with the target in the future. In evolutionary models, third-party punishment is inherently costly (Raihani, Thornton, & Bshary, 2012). To operationalize this idea, lab experiments on costly punishment typically impose a direct cost on the choice to punish.

Therefore, each situation in which an individual must decide whether to punish, or not, can be characterised by the expected consequences of punishing for the punisher, the social good, and the target. Punishers consider all these consequences when deciding whether to punish (Wiessner, 2005; Twardawski, Tang, & Hilbig, 2020; Berg, Kitayama, & Kross, 2021; Marshall, Yudkin, & Crockett, 2021); we refer to a punisher who considers these three consequences as the "Base punisher". However, different individuals may vary in how much they care about each of these expected consequences, and how strongly they weigh those in their decisions.

We formalize these ideas, using the following definition for the base punisher's expected utility over each action 'a'.

$$U_{BP}(a) = \alpha_{self} U_{self}(a) + \alpha_{social} U_{social}(a) + \alpha_{target} U_{target}(a) \quad (1)$$

where U_x represent the subjective utility that the punisher assigns to each consequence. Each α_x is the weight the individual places on the corresponding utility component, U_x. Therefore, the αs represent the hidden motives and values of the base punisher. For instance, a high α_{self} represents an individual who cares a lot about direct consequences (costs and benefits) for herself. Hereafter, we use $\boldsymbol{U}$ to denote the set of $\{U_{self}, U_{social}, U_{target}\}$, and $\boldsymbol{\alpha}$ to denote $\{\alpha_{self}, \alpha_{social}, \alpha_{target}\}$.

Actions are selected using a softmax decision rule:

$$P(a|\boldsymbol{\alpha}, \boldsymbol{U}) \propto exp(\beta U_{BP}(a)) \quad (2)$$

Communicative punisher

Compared to the base punisher, the "Communicative Punisher" has an additional preference over how their motives and values are perceived by a relevant audience. This desired *impression* incurs an additional utility term, i.e., $U_{reputation}$, which is defined for each action 'a', as the probability that the action will evoke the desired impression. Therefore,

$$U_{CP}(a) = U_{BP}(a) + \alpha_{reputation} U_{reputation}(a) \quad (3)$$

$$U_{reputation}(a) = P(impression|a)$$

This desired impression is what the punisher wants the audience to infer about the values of their $\boldsymbol{\alpha}$. Therefore, to estimate $U_{reputation}$, the punisher needs to know how the beliefs of the audience about the $\boldsymbol{\alpha}$ change, after observing an action by the punisher.

For this, the communicative punisher uses their mental model of an audience, who in turn, recursively represents a base punisher. This audience model will update its impressions of the punisher's $\boldsymbol{\alpha}$ by performing Bayesian inverse planning, given the punisher's action and the audience's prior belief about the preferences of base punishers, $P(\boldsymbol{\alpha})$.

$$P(\boldsymbol{\alpha}|a, \boldsymbol{U}) \propto P(a|\boldsymbol{\alpha}, \boldsymbol{U})\, P(\boldsymbol{\alpha}) \quad (4)$$

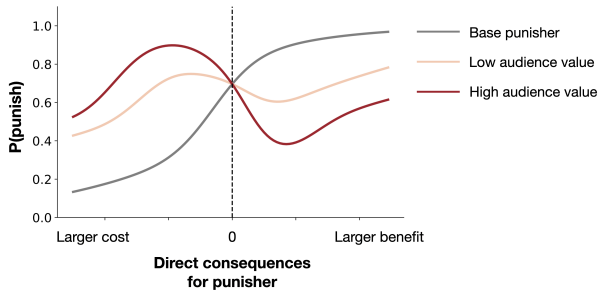

Figure 2: How selfish costs and rewards affect punishment. For the base punisher, punishment is less likely when it is costly and more likely when it is rewarded. For communicative punishers who want to avoid creating an impression of overly selfish motives, this relation is reversed: costly punishment is more likely and rewarded punishment is less likely. The more the punisher values the audience, the stronger this effect.

where $P(a|\boldsymbol{\alpha},\boldsymbol{U})$ is the base punisher's policy, derived from equation (2). Note that here we assume that the punisher and audience have common knowledge of $\boldsymbol{U}$. More generally, the audience could be uncertain about the $\boldsymbol{U}$, could have different beliefs about $\boldsymbol{U}$ from the punisher, and could update beliefs about $\boldsymbol{U}$ from observing the punisher's actions. Here, we restrict the audience inferences to $\boldsymbol{\alpha}$, the weight that the punisher places on the commonly known $\boldsymbol{U}$. The posterior belief over $\boldsymbol{\alpha}$ can then be flexibly used to estimate judgements for any impression that the punisher prefers.

$$P(impression|a) = \sum_{\boldsymbol{\alpha}} P(impression|\boldsymbol{\alpha})P(\boldsymbol{\alpha}|a,\boldsymbol{U}) \qquad (5)$$

Desiring to be perceived as unselfish

The RCSA framework is highly flexible: potential punishers may want to create many possible impressions. By contrast, evolutionary and game-theoretic models typically assume that agents are mainly motivated to communicate that they are not selfish, and so can be trusted not to exploit future interaction partners. Therefore, we began by modeling potential punishers who want to be seen as unselfish.

Rationally choosing actions in order to be seen as unselfish creates a distinctive signature in decisions (Figure 2), that could be confirmed experimentally. Holding everything else constant, these communicative agents will be more likely to choose an action as its cost increases, and less likely to choose the action as its reward increases. The strength of this effect depends on how much the person values the opinion of the audience.

Experimental findings

We conducted a literature review to identify existing experiments that manipulated the direct selfish cost or reward, or the audience, for third-party punishment. We identified three recently published papers that contain relevant experimental manipulations (J. J. Jordan & Rand, 2020; Batistoni, Barclay, & Raihani, 2022; Rai, in press).

Costly third-party punishment is chosen to communicate unselfishness

As shown in Figure 2, people who want to be seen as unselfish by a valued audience may be *more* likely to punish, when punishment is costly, than when punishment is not costly. However, we were not able to identify any experiments that hold everything else constant, and manipulate only the cost of punishment. Thus, we focused instead on the predicted effect of varying the value of the audience. We selected two recent papers that measured the effect of a relevant audience on choices to punish, when punishment is costly, in anonymous multi-player financial interactions.

In Batistoni et al. (2022), third-party punishers were endowed with \$0.50 and could decide to invest any amount between \$0.00 and \$0.45 to punish a target who makes an unfair allocation to another stranger, in a dictator game. More costly punishments imposed harsher costs on the target of punishment. Then, the third-party punisher played as a trustee in a trust game with another participant, the investor. Trustees benefit from being seen as unlikely to exploit their investors. The future investor either observed the punishment choice ("Observed"), or did not ("Private/Unspecified"). People were more likely to punish, and paid larger costs to execute harsher punishments, when observed by their future investors (See figure 3A).

In J. J. Jordan and Rand (2020), experiments 9 and 10, third-party punishers were given the opportunity to punish a target who had behaved selfishly toward a stranger. To manipulate the audience value, as above, half of the punishers were observed by the investor for a subsequent trust game ("Observed"); the other half of punishers did not play a subsequent trust game ("Private/Unspecified"). Again, people were more likely to punish when observed by a relevant audience (Figure 3B, "P-only"). The key innovation of J. J. Jordan and Rand (2020) was that some third-party punishers were given an opportunity to first offer costly help to another anonymous target, and some were not. We refer to these two conditions as "Punishment-only" and "Help+Punishment". Costly helping could serve as a sufficient opportunity to communicate unselfishness, reducing the reputation benefits of subsequent punishment choices. As predicted, when observed by a relevant audience, people were less likely to punish after being given the opportunity to help (Figure 3B, "H+P").

Beneficial punishment is avoided to communicate unselfishness

Perhaps the most striking effect of wanting to avoid being seen as selfish, in Figure 2, is that increasing the direct reward for punishing actually decreases the probability of choosing to punish. Direct rewards for third party punishment are rarely studied, but one recent paper directly tested this prediction.

In Rai (in press), Study 1, third-party punishers could decide to punish a target who makes an unfair allocation to another stranger, in a dictator game. Participants were randomly assigned to be offered either no payment, or a small payment, if they chose to punish. People were less likely to punish if punishing received a small reward (Figure 3C). In Rai (in press), Study 3a, participants were asked to imagine an opportunity to punish a co-worker for insider trading. Participants were told that they would receive either no rewards, a small reward or a very large reward for their punitive action. People were less willing to punish if offered a small reward than if offered no reward. Willingness to punish recovered in response to very large rewards (Figure 3D).

In the next section, we demonstrate how RCSA can be used as a quantitative framework to capture these findings.

Model specification

To build a concrete RCSA model, we need to determine the utilities of each action, specify the configuration of parameters, and specify the internal model of the audience as well as formalizing the communicative goal of punisher. The specific values of utilities and parameters used for each dataset are reported in table 1. Note that we did not systematically search for the set of parameters that best fit the data quantitatively, as the focus of the current paper is not on best fitting the data, but on exploring RCSA as a unified computational framework to synthesize patterns of human behaviour.

Utilities Existing experiments have mostly operationalized punishments as monetary decisions with deterministic monetary outcomes for the target and the punisher. Therefore, we assume that the subjective utilities, U_{self} and U_{target}, change monotonically with the monetary costs and benefits that are determined by the experimental design. However, these subjective utilities are not necessarily identical to the objective costs and benefits. For example, we considered the negative utility of harming the target to be larger than the positive utility of helping the target with the same amount of money, supported by existing work showing that people are averse to harming others (Cushman, Gray, Gaffey, & Mendes, 2012). Specifying U_{social} is more challenging, because the effects of punishment on both specific and general deterrence are disputed. For now, we assume that punishers, and audiences, share an expectation that punishment will cause some social good. We therefore chose positive values for U_{social}, within the same order of magnitude as U_{target}.

Variability in punisher's values Individuals can vary in how much they care about their own outcomes, α_{self}, the social good, α_{social} and the specific target of punishment, α_{target}. We simulate a population of individuals, in which α_{self} and α_{social} are independent and $\alpha_{self} \sim \text{Exponential}(\lambda)$, with $\lambda = \frac{1}{3}$, and $\alpha_{social} \sim \text{Uniform}(0,10)$. No other parameters varied across individuals. In all of the experiments considered here, the punisher and the target of punishment are anonymous strangers, so we set a fixed small value for α_{target}. For comparison to experimental results, which measure the probability of punishing in a population of participants, model results are plotted as the average behaviour of this population.

Audience inferences We simplified the audience's inference to joint inference of α_{self} and α_{social}, given the observed actions. The audience knows that the target of punishment is an anonymous stranger, and so does not make inferences about α_{target}. Although the true values of α_{self} and α_{social} in the population vary continuously, we assumed that punisher's internal model of the audience's inference is relatively coarse-grained. Punishers want to be seen as "selfless enough", rather than as "completely selfless". So, in the internal model of the audience, the weights, i.e., α, on selfish and social outcomes can take either a relatively small or large value. The audience's prior over these types is uninformative, assuming a uniform distribution in the population of all types.

We formalized being perceived as unselfish as having a low α_{self}, plugging in the following definition into equation 5:

$$P(impression|\boldsymbol{\alpha}) = \begin{cases} 1 & small\ \ \alpha_{self} \\ 0 & large\ \ \alpha_{self} \end{cases} \tag{6}$$

Finally, we varied how much weight the potential punisher puts on the audience's inferences. A potential punisher will likely care most about the audience's inferences when the audience's opinions are consequential for her future. In reality, highly valued audiences are likely to be high status in-group members. In typical experiments, the most relevant audiences are anonymous strangers who will interact with the participant in a subsequent financial exchange. By contrast, in the control conditions, no audience is mentioned at all. Yet, people act as if considering some audience, even when actions are described as anonymous (Dana, Cain, & Dawes, 2006). Thus we model the experimental conditions as a high audience value, when a relevant observer was specified by the experiment, and a low audience value, when no observer was specified.

Model Results

We used RCSA to model the effect of being observed on costly punishment (J. J. Jordan & Rand, 2020; Batistoni et al., 2022), the effect of a parallel situation to communicate unselfishness (J. J. Jordan & Rand, 2020), and the effect of directly rewarding punishment (Rai, in press). The code for the model simulations and analysis of experimental data are available at `https://github.com/sradkani/punishment-rcsa-cogsci22`.

Batistoni et al. (2022) To model this experiment, we considered a two-stage hierarchical decision process. In the first stage, the potential punisher faces a binary decision of whether or not to punish. In the second stage, only those who chose to punish in the first stage decide between 9 options, i.e., what proportion of their endowment to invest in punishment (from 0.1 to 1 in steps of 0.1). The cost of punishment

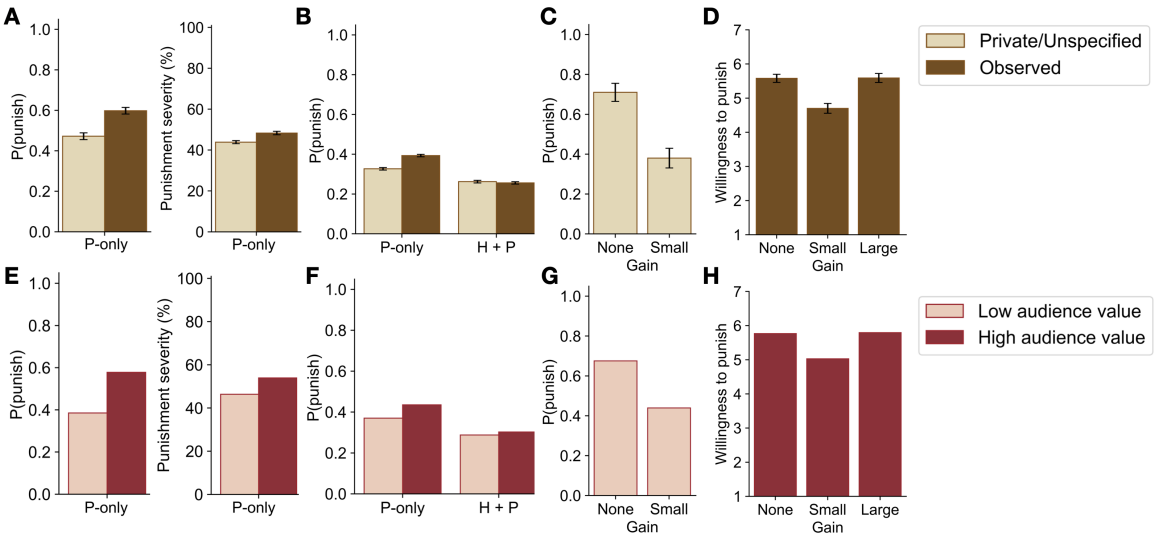

Figure 3: Existing data, top row: Costly punishment in (**A**) Batistoni et al. (2022) and (**B**) J. J. Jordan and Rand (2020). Rewarded punishment in Rai (in press) (**C**) study 1; and (**D**) study 3a. Corresponding modeling results, bottom row: **E:H**. "P-only": participants could choose to punish, or not. "H+P": before the punishment decision, participants could choose to help. Unobserved choices were modeled as "Low audience value" and choices observed by a relevant audience were modeled as "High audience value".

(U_{self}), and the imposed cost on the target (U_{target}) scale linearly with the amount of investment. U_{social} starts low and saturates at high values of endowment, following a sigmoid function. For stage 1, the utilities of the "Punish" decision are obtained by taking the mean of utilities for the 9 severity options.

Figure 3E shows model simulations for communicative punishers, in settings where the audience value ($\alpha_{reputation}$) is high or low. Similar to the experimental data, in front of a more valued audience (i.e., higher $\alpha_{reputation}$), communicative punishers are more likely to punish, and use harsher punishments.

J. J. Jordan and Rand (2020) In the "Punishment-only" condition, we modeled a potential punisher who faces a binary decision of whether or not to punish. To model the "Help+Punishment" condition, we considered communicative punishers who had made a decision whether to help or not, before having the opportunity to punish. Helping decisions in this experiment have similar consequences to punishment. Helping benefits the target, U_{target}, achieves a future social benefit, U_{social}, and has direct costs for the helper, U_{self}. Therefore, using the same RCSA computations, this punisher can estimate audience's posterior belief, after observing a person choose costly helping. The punisher then uses this posterior as the new prior in equation 4, in order to estimate $U_{reputation}$ of punishment and decide whether to punish. To find the probability of punishing in the "Help+Punishment" condition, we averaged the behavior of potential punishers who had helped and not helped before, weighted by their prevalence in the empirical sample (65% helped in "Private/Unspecified" and 83% helped in "Observed" condition).

Figure 3F shows model simulations for communicative punishers, when acting in front of audiences with high ("Observed") or low ("Private/Unspecified") value to the punisher, in the "Punishment-only" and "Help+Punishment" conditions. First, in front of the more valued audience, communicative punishers are more likely to punish. Second, the communicative value of punishment decreases when punishment is preceded by a more informative action. As a result, punishment is less likely in "H + P" compared to "P-only" conditions, particularly when in front of a highly valued audience.

Rai (in press) We modeled a potential punisher who has been offered zero or small U_{self} for punishing (see table 1). Because the audience was unspecified, we used a small $\alpha_{reputation}$. Consistent with the experimental results, the model predicts that communicative punishers avoid moderately rewarded punishment (Figure 3G). Then, with a much more valued audience ("co-workers", in the experiment, modeled as a large $\alpha_{reputation}$), we estimated willingness to punish when offered no reward, a small U_{self}, or a very large U_{self} for punishing. Again, similar to the data, the model predicts less willingness to punish when punishment is slightly rewarded compared to when it has no direct benefits for the punisher; however, willingness to punish recovers for larger benefits, when U_{self} dominates the effect of $U_{reputation}$ (Figure 3H).

Table 1: Parameters for modeling existing datasets. In Batistoni et al. (2022), participants chose the fraction of their endowment to invest in punishment (p, ranging from 0.1 to 1). In J. J. Jordan and Rand (2020), some participants had a chance to help before choosing whether to punish. In Rai (in press), participants were offered zero, small or large rewards for punishing. All utilities for decisions of "Not-punish" and "Not-help" are set to zero. See text for more details.

Dataset	U_{self}	U_{social}	U_{target}	α_{self}	α_{social}	α_{target}	$\alpha_{reputation}$	β
Batistoni et al., 2022	$-45p$	$\frac{10}{1+e^{10(0.5-p)}}$	$-100p$	Small: 2 Large: 8	Small: 2 Large: 8	2	Small: 250 Large: 350	0.02
Jordan & Rand, 2020	Punish: -5 Help: -20	10	Punish: -50 Help: 15					
Rai, 2022 (Study 1)	None: 0 Small: 15	10	-5				250	
Rai, 2022 (Study 3a)	None: 0 Small: 5 Large: 500	100	-100				1200	

Discussion

People choose to punish others not only to induce the direct consequences of punishment, but also to evoke specific desired inferences about the punisher's values in the observers. We introduced Rational Communicative Social Action (RCSA) as a computational framework to model punishment decisions. In the RCSA model, a potential punisher uses a recursive model of the audience to anticipate their inferences about the punisher's underlying values and motives. The punisher then rationally balances a desire for this inference against the direct consequences of punishment for the target, the punisher and the social good. We illustrated how empirically observed patterns of human punitive decisions can be captured using this framework.

RCSA is a model-based planning algorithm (Ho et al., 2021). That is, RCSA proposes that people make sophisticated, flexible, context-sensitive plans to punish, using utilities expressed over recursive inferences about observers' perceptions of the punisher. We are not suggesting that people *always* plan to punish; habits or heuristics may offer cognitively cheaper ways to make punitive choices in familiar or repetitive situations (Dezfouli & Balleine, 2012). People do use heuristics in punishment and more broadly prosocial decision making (Rand et al., 2014; J. Jordan & Kteily, 2020).

However, we specifically hypothesize that humans can and do use rich, recursive model-based plans to choose whether or not to punish in new, unfamiliar or particularly consequential situations (Kool, Cushman, & Gershman, 2016). Existing experiments provide initial evidence for this flexibility: for example, it is unlikely that people have formed habits of punishing less after being observed in costly helping, or of avoiding rewarded third-party punishment.

The RSCA model of punishment is both general and flexible. In the current work, we have only explored the effects of one desired impression on punitive decisions: the desire to appear unselfish. In some situations, potential punishers are motivated to communicate to the audience that they will not exploit future interaction partners. However, we find it plausible that depending on the types of future interactions the punisher is going to have with the audience, she may pursue other desired social perceptions. For example, punishers may wish to communicate that they care about the target, share the audience's values (J. Jordan & Kteily, 2020), or are dominant or fearsome (Raihani & Bshary, 2015). RCSA can be used to formalize such desires using a similar logic used here, for example, by specifying a desire on the inferred value of α_{social} and α_{target}.

Preferences for how the punisher is socially perceived could interact in interesting ways with other communicative goals of punishment, i.e., communicating about the transgression itself (Sarin et al., 2021). Indeed, these two communicative goals may in some cases conflict. For example, harsh punishments could communicate more extreme disapproval of the transgression, while at the same time risk communicating that the punisher is selfish or callous. The RCSA model could be further extended to integrate both of these types of communicative goals for punishment.

Overall, the goal of the RCSA models is to capture the cognitive process that humans actually use when making punitive decisions in real life. If successful, these models could tell us something about the internal computations and representations that people actually entertain while making choices. The models show how people could make punitive choices that rationally incorporate communicative intentions towards a specific audience, and their trade-offs with other instrumental goals of punishing. These models can also be used to generate quantitative predictions for new experiments that more systematically explore the range of communicative goals people can pursue by punishing. For instance, a stronger quantitative test of the framework would ideally test all of the predictions implicit in the curves of Figure 2, by varying one parameter at a time while holding all other utility components constant. We hope to conduct such tests in future studies.

Acknowledgments

This work was supported by the Patrick J. McGovern Foundation grant, and Mathworks Fellowship. We thank J. J. Jordan and Rand (2020) and Batistoni et al. (2022) for making their data available.

References

Asaba, M., & Gweon, H. (2019, Dec). *Young children infer and manage what others think of the self.* PsyArXiv. Retrieved from psyarxiv.com/yxhv5 doi: 10.31234/osf.io/yxhv5

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.

Berg, M. K., Kitayama, S., & Kross, E. (2021). How relationships bias moral reasoning: Neural and self-report evidence. *Journal of Experimental Social Psychology*, *95*, 104156.

Bregant, J., Shaw, A., & Kinzler, K. D. (2016). Intuitive jurisprudence: Early reasoning about the functions of punishment. *Journal of Empirical Legal Studies*, *13*(4), 693–717.

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

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: the aversion to harmful action. *Emotion*, *12*(1), 2.

Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and human decision Processes*, *100*(2), 193–201.

Dezfouli, A., & Balleine, B. W. (2012). Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, *35*(7), 1036–1051.

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*(1), 201–224.

Feinberg, J. (1965). The expressive function of punishment. *The Monist*, *49*(3), 397–423.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., & Griffiths, T. L. (2021). Control of mental representations in human planning. *arXiv preprint arXiv:2105.06948*.

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2019). People teach with rewards and punishments as communication, not reinforcements. *Journal of Experimental Psychology: General*, *148*(3), 520.

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.

Kleiman-Weiner, M., Shaw, A., & Tenenbaum, J. (2017). Constructing social preferences from anticipated judgments: When impartial inequity is fair and why? In *Cogsci.*

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When does model-based control pay off? *PLoS computational biology*, *12*(8), e1005090.

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human behavior*, *28*(2), 75–84.

Marshall, J., Yudkin, D. A., & Crockett, M. J. (2021). Children punish third parties to satisfy both consequentialist and retributive motives. *Nature Human Behaviour*, *5*(3), 361–368.

Mulder, L. B. (2018). When sanctions convey moral norms. *European Journal of Law and Economics*, *46*(3), 331–342.

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

Nelissen, R. M. (2008). The price you pay: Cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior*, *29*(4), 242–248.

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.

Primoratz, I. (1989). Punishment as language. *Philosophy*, *64*(248), 187–205.

Rai, T. S. (in press). Material benefits crowd out moralistic punishment. *Psychological Science*.

Raihani, N. J., & Bshary, R. (2015). The reputation of punishers. *Trends in ecology & evolution*, *30*(2), 98–103.

Raihani, N. J., Thornton, A., & Bshary, R. (2012). Punishment and cooperation in nature. *Trends in ecology & evolution*, *27*(5), 288–295.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature communications*, *5*(1), 1–12.

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.

Sunstein, C. R. (1996). On the expressive function of law. *University of Pennsylvania law review*, *144*(5), 2021–2053.

Sunstein, C. R., Kahneman, D., & Schkade, D. (1997). Assessing punitive damages (with notes on cognition and valuation in law). *Yale Lj*, *107*, 2071.

Twardawski, M., Tang, K. T., & Hilbig, B. E. (2020). Is it all about retribution? the flexibility of punishment goals. *Social justice research*, 195–218.

Wiessner, P. (2005). Norm enforcement among the ju/'hoansi bushmen. *Human Nature*, *16*(2), 115–145.

Yoon, E. J., MacDonald, K., Asaba, M., Gweon, H., & Frank, M. C. (2018). Balancing informational and social goals in active learning. In *Cogsci.*

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2020). Polite speech emerges from competing social goals. *Open Mind*, *4*, 71–87.