A computational model of strategic punishment in divided societies

Hector X. Martinez (hectorxm@mit.edu)[1], Rebecca Saxe (saxe@mit.edu)[2] &
Setayesh Radkani (setayesh_radkani@brown.edu)[3]
[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA
[2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139 USA
[3]Department of Cognitive and Psychological Sciences, Brown University, Providence, RI 02912 USA

Abstract

In divided societies, authorities often use punishment to establish shared norms while trying to maintain or signal their legitimacy. When facing a polarized audience, these goals may often be at odds. We extend the Rational Communicative Social Action (RCSA) framework (Radkani et al., 2022) to formally model an authority's strategic decision-making when using public punitive actions to pursue these goals. We distinguish between a communicative authority who aims to shape the audience's moral beliefs, and a reputation-aware authority who optimizes the audience's assessment of their own character. By simulating these agents against polarized audiences, we characterized the tradeoffs and dynamics of strategic punishment when facing diverse audiences with various forms and levels of polarization, which serves to generate systematic predictions for future experiments.

Keywords: Polarization; Punishment; Bayesian inference; Theory of mind; Social signaling

Introduction

In 2023, professor Erika López Prater included 14th century images of the prophet Muhammad in her world art class at Hamline college; after students complained that this constituted Islamophobia, professor Prater's contract was not renewed. The President of Hamline college was subjected to nation-wide scrutiny after this event, as diverse audiences debated both the initial event (was showing the image disrespectful to Muslim students?) and the punitive response (was President Miller biased against academic freedom?). This example illustrates a key challenge facing authorities: how to use punishment to sustain community values and maintain legitimacy, when the community may have diverse and even polarized beliefs about both the target act and the authority's motives.

Evolutionary and game-theoretic models often assume punishment has fixed positive consequences, such as enforcing shared norms of cooperation in the society (Fehr and Gächter, 2002) or increasing the perceived trustworthiness of the punisher (Raihani and Bshary, 2015; Santos et al., 2011), which in turn motivates people to punish. While successful in specific contexts (e.g., Jordan et al., 2016), these models fail to capture the context-dependence of human punishment. First, punishment has variable consequences (Radkani et al., 2025): it may successfully signal wrongness (L. B. Mulder et al., 2009; Nosenzo et al., 2024) or fail to do so (Depoorter and Vanneste, 2005; Nosenzo et al., 2024; Verboon and van Dijke, 2011; Xiao, 2013); it may enhance trust (Dhaliwal et al., 2020; Jordan et al., 2016) or lead to perceptions of illegitimacy (Eriksson et al., 2017; Sun et al., 2023). These outcomes depend on punishment severity (L. Mulder, 2018; Salcedo and Jimenez-Leal, 2024), the punisher's relationship to the target (Gordon and Lea, 2016; Marshall et al., 2020), and the audience's pre-existing beliefs about the wrongness of the target act and the authority's motives (L. B. Mulder et al., 2009; Nosenzo et al., 2024; Radkani et al., 2025). Second, when contemplating whether to punish a norm violation, people anticipate these variable consequences and adjust their punitive response accordingly, depending on the key features of the punitive context such as the audience's ideology (Jordan and Kteily, 2025), group memberships (Ashokkumar et al., 2019), and the material consequences of punishment (Rai, 2022; Sarin et al., 2021).

In this project, we developed a cognitively-inspired computational framework that formalizes how authorities anticipate the context-dependent consequences of punishment. Building on the Rational Communicative Social Action framework (RCSA; Radkani et al., 2022) and a validated model of audience inferences from punishment (Radkani et al., 2025), we modeled how authorities reason about and make punitive decisions through recursive Bayesian inference. Next, we used model simulations to study how audience polarization shapes authorities' punitive decisions, generating systematic cognition-inspired predictions that can be tested in future experiments.

RCSA: A cognitive framework for modeling strategic punishment

Punishment is not merely a deterrent cost; it is a social action. From observing an authority's punitive choices, people make inferences about the underlying motives and character of the authority (Radkani et al., 2025), which in turn shapes their perception of the authority's legitimacy (Saxe, 2022). At the same time, observers update their beliefs about the wrongness of the target act and learn about the norms of the society (L. Mulder, 2018; Radkani et al., 2025; Sarin et al., 2021). Potential punishers, in turn, anticipate the audience's inferences (Rai, 2022; Sarin et al., 2021) and use punishment to both strategically shape these perceptions (Rai, 2022; Tsai, 2021), and to communicate moral values and establish norms (Funk et al., 2014; Ho et al., 2019; Sarin et al., 2021).

To formally capture strategic punitive decision-making,

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

Radkani et al. (2022) developed the RCSA model of punishment inspired by recursive models of pragmatic reasoning in language (Goodman and Frank, 2016). In these models, a speaker plans their utterances to communicate hidden meaning to a listener by reasoning over a mental model of how the listener interprets the speaker's words by rationally reasoning over a mental model of a literal speaker. Similarly, in RCSA, the authority plans their punishment by recursive reasoning over a mental model of how the audience interprets the authority's punishment by rationally reasoning over a mental model of a non-strategic naive authority.

However, the existing RCSA model has two important limitations. First, it focuses only on authorities' reputational concerns and does not capture the didactic goal of communicating social norms. Second, mirroring much of the empirical literature, it considers only a narrow subset of reputation concerns, namely, the desire to appear unselfish. We extend the RCSA framework and formalize how authorities anticipate and plan to shape audience inferences about both norms and other core dimensions of legitimacy in punitive contexts, including justice motive and impartiality (Tyler, 2006; Saxe, 2022).

Extending the RCSA framework

We embed an empirically validated cognitive model of the audience's joint inferences of moral norms and authority's motives (Radkani et al., 2025) in the planning model of a strategic authority. The audience's inferences are in turn driven by inverse planning over a mental model of a naive authority.

Naive authority

A naive authority makes decisions solely based on their intrinsic motives, without considering the audience inferences. This agent chooses a punitive action a from a discrete set of options $A = \{\text{None}, \text{Mild}, \text{Harsh}\}$ to maximize an intrinsic utility function $U_{\text{int}}(a)$.

Inspired by the literature on punitive motivations, we modeled the naive authority as pursuing three sources of utility. First, an authority cares about the direct consequences of their decision for themselves ($U_{\text{self}}(a)$; Fehr and Gächter, 2002, Rai, 2022). Second, they care about the consequences for the target, e.g., how much punitive harm they impose on them ($U_{\text{target}}(a)$; Raihani and Bshary, 2019). Third, they care about responding proportionately to the transgression ($U_{\text{justice}}(a)$), by balancing the severity of the punitive action against the moral wrongness of the target act ($m \in [0, 1]$) (Carlsmith et al., 2002; Sznycer and Patrick, 2020). Here, we assume the severity of action a is equal to the consequences it imposes on the target (i.e., $U_{\text{target}}(a)$).

$$U_{\text{justice}}(a) = -|m + U_{\text{target}}(a)| \tag{1}$$

The total intrinsic utility of an action is the sum of these three utilities, weighted by the naive authority's selfish motive ($\alpha_{\text{self}} \in [0, 1]$), bias towards the target ($\alpha_{\text{target}} \in [-0.5, 0.5]$, with negative values for bias against and positive values for bias in favor, and justice motive ($\alpha_{\text{justice}} \in [0, 1]$).

$$U_{\text{int}}(a) = \alpha_{\text{justice}} U_{\text{justice}}(a) + \alpha_{\text{target}} U_{\text{target}}(a) + \alpha_{\text{self}} U_{\text{self}}(a) \tag{2}$$

The naive authority selects an action according to a softmax policy, given their belief about wrongness of the act (m) and their set of motives ($\boldsymbol{\alpha}$).

$$P(a|m, \boldsymbol{\alpha}) \propto \exp(\beta \cdot U_{\text{int}}(a)) \tag{3}$$

In what follows, we set $U_{\text{self}}(a) = 0$ for all a, focusing on settings where selfish gains are not a salient concern. This allows us to isolate the core dynamics of audience beliefs and authorities' planning in contexts where group affiliation generates divergent moral values and asymmetric beliefs about authority legitimacy, as in many real-world political and social conflicts.

Strategic authority

Relative to the naive authority, the strategic authority additionally cares about shaping the audience's beliefs about the wrongness of the act (i.e., communicating norms) and about the authority's motives (i.e., managing reputation). However, because the authority does not have direct access to the audience's actual beliefs or belief updates, it must rely on an intuitive mental model of the audience. Specifically, the authority assumes an audience with particular prior beliefs and uses this mental model to anticipate how those beliefs will be updated in response to its punitive decisions.

Recursive reasoning to anticipate audience inferences To model how authorities anticipate the consequence of potential actions for the audience's beliefs, we embed within the authority a validated model of the audience's inferences (Radkani et al., 2025). In this model, the audience interprets punishment through inverse planning over a mental model of a naive authority, treating the authority as a rational agent who chooses actions to maximize the intrinsic utility function $U_{int}(\cdot)$ described above (Equation 2). An observer maintains a joint probability distribution $P(m, b, j)$ over the wrongness of the act, and the authority's bias and commitment to justice. Upon observing an action a_{obs}, the audience updates their prior beliefs via Bayesian inference. The likelihood of observing a specific action a under the hypothesis (m, b, j) is given by the softmax policy in Equation 3.

$$P(m, b, j|a_{\text{obs}}) \propto P(a_{\text{obs}}|m, b, j) P(m, b, j) \tag{4}$$

The inverse planning procedure for this belief update is detailed in Algorithm 1, implemented using the JAX package in Python (Frostig et al., 2018). The inference is implemented numerically over a discretized grid for the continuous variables m, b and j.

The authority then uses this joint posterior distribution $P(m, b, j|a)$ to compute the strategic utility of each action a, given their reputational and/or communicative desire.

Algorithm 1 Anticipate audience belief update through inverse planning

Require: Prior belief distribution $P_0(m, b, j)$
Require: Observed action $a_{\text{obs}} \in \{\text{None, Mild, Harsh}\}$ with corresponding U_{target} values
Require: Rationality parameter β_{obs}, Hypothesis grid G
Step 1: Compute Likelihoods
for each state tuple (m, b, j) in G **do**
  Calculate intrinsic utility vector U_{int} for all actions a':
    $U_{\text{int}}(a') \leftarrow -j|m + U_{\text{target}}(a')| + bU_{\text{target}}(a')$
  Compute probability of observed action (Softmax):
    $P(a_{\text{obs}} \mid m, b, j) \leftarrow \frac{\exp(\beta_{\text{obs}} U_{\text{int}}(a_{\text{obs}}))}{\sum_{a' \in A} \exp(\beta_{\text{obs}} U_{\text{int}}(a'))}$
end for
Step 2: Bayesian Update
$P_{\text{unnorm}} \leftarrow P(a_{\text{obs}} \mid m, b, j) \times P_0(m, b, j)$ {Posterior belief}
$P_1(m, b, j) \leftarrow P_{\text{unnorm}} / \sum P_{\text{unnorm}}$ {Normalize}
return $P_1(m, b, j)$

Here, we assume that the authority faces a diverse audience composed of two homogeneous groups, which we label the in-group and out-group. This terminology is mainly for convenience and does not imply that the dynamics we study are specific to group membership.

Reputational desire Authorities may use punishment to shape audience perceptions of their legitimacy (Tsai, 2021; Rai, 2022); and common intuitions about legitimacy emphasize benevolence and impartiality (Tyler, 2006; Saxe, 2022). In punitive contexts, we formalize this desire as a reputational utility that increases when the authority is perceived as highly motivated by justice and as impartial (i.e., not biased in favor or against the target).

$$U_{\text{rep}}(a) = k_{\text{j\_in}}\mathbb{E}[j_{\text{in}}|a] + k_{\text{j\_out}}\mathbb{E}[j_{\text{out}}|a] \qquad (5)$$
$$-k_{\text{b\_in}}|\mathbb{E}[b_{\text{in}}|a]| - k_{\text{b\_out}}|\mathbb{E}[b_{\text{out}}|a]|$$

, where $(j_{\{\text{in, out}\}}|a)$ denotes the marginalized posterior beliefs of the in-group and out-group about the authority's justice motive given action a, and $(b_{\{\text{in, out}\}}|a)$ denotes the corresponding posterior beliefs about the authority's bias. Both positive (bias in favor) and negative (bias against) values of b incur a negative reputational utility. $\mathbb{E}$ is the expected value of these belief distributions, and $k_{\{\text{j/b}\}\_\{\text{in/out}\}}$ are fixed constants determining the weight the authority assigns to the audience beliefs about its justice and bias.

Communicative desire Punishment is often used to communicate the wrongness of the target act and to establish shared norms of behavior (Funk et al., 2014; Ho et al., 2019; Sarin et al., 2021). We formalize this desire as a communicative utility that increases when the audience's moral beliefs $(P(m_{\{in/out\}}|a))$ become more aligned with the authority's own moral judgment (m^*).

$$U_{\text{comm}}(a) = -k_{\text{m\_in}}D(P(m_{in}|a), m^*) -$$
$$k_{\text{m\_out}}D(P(m_{out}|a), m^*) \qquad (6)$$

To quantify the degree of misalignment between the audience's beliefs and the authority's moral value, we use the Wasserstein-1 (Earth Mover's) distance to a point mass at m^*:

$$D(P, m^*) := W_1(P, \delta_{m^*}) = \sum_m P(m)|m - m^*| \qquad (7)$$

The Wasserstein-1 distance is sensitive to how far beliefs lie from the authority's target belief, and therefore gives greater weight to shifting audiences who are farther from that target.

A strategic authority selects an action according to a softmax policy with the total utility of an action a given as:

$$U_{\text{strategic}}(a) := U_{\text{int}}(a) + \alpha_{\text{rep}}U_{\text{rep}}(a) + \alpha_{\text{comm}}U_{\text{comm}}(a) \quad (8)$$

, where α_{rep} and α_{comm} denote the extent to which the authority cares about managing their reputation and communicating norms, respectively.

Simulations

We used our extended RCSA framework to systematically study the punitive policy of strategic authorities when facing diverse audiences with different forms and levels of polarization. In particular, we simulated two types of polarized societies that authorities are likely to face: a society where people are polarized in their moral values, and a society where people vary in their level of trust toward the authority. We used our model to simulate how authorities evaluate the reputational and communicative utility of each punitive response (i.e., harsh punishment, mild punishment, and inaction), under different forms and levels of audience polarization. We also varied the authority's own moral conviction (m^*) to study its effect on these utilities. Across all simulations, we set $U_{\text{target}}(\text{None}) = 0$, $U_{\text{target}}(\text{Mild}) = -0.5$, $U_{\text{target}}(\text{Harsh}) = -1$, $\beta = \beta_{\text{obs}} = 10$, $k_{\{j/b\}\_\{\text{in, out}\}} = 2$, and $k_{\text{m\_}\{\text{in, out}\}} = 1$. The code for model simulations is available at https://github.com/hectorxm25/strategic-punisher.

Polarization in moral beliefs

In this setting, the audience consists of two groups who are polarized in their prior beliefs about the wrongness of the target act: one group a priori believes the act is wrong, while the other group apriori believes the act is relatively innocuous. We express the degree of social fracture with a polarization parameter $p \in [0, 1]$, which controls the separation between the two groups' wrongness priors. Audience prior beliefs are modeled as beta distributions. The groups' mean beliefs are given by $\mu_{\text{in}} = 0.5 + 0.45p$ and $\mu_{\text{out}} = 0.5 - 0.45p$. Thus, when $p = 0$, both groups hold shared beliefs about wrongness of the act ($\mu_{in/out} = 0.5$); when $p = 1$, they are strongly polarized ($\mu_{in} = 0.95$ and $\mu_{out} = 0.05$). Only the means vary across simulations; the variance is held fixed at approximately 0.01. Both groups share uncertain beliefs about the authority's justice motive and bias, modeled as uniform distributions over these two variables.

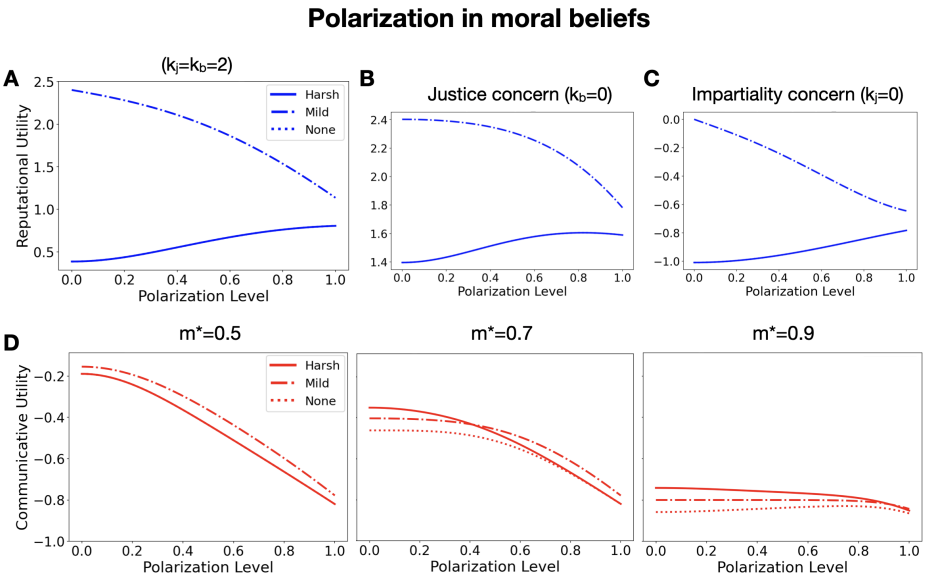

Figure 1: Reputational and communicative utility of punishment when facing audiences polarized in their moral beliefs. A) overall reputational utility of harsh punishment, mild punishment, and inaction as a function of the polarization parameter p that controls the distance between the two groups' beliefs about wrongness; B) reputational utility associated with being perceived as just, with no weight placed on impartiality concerns ($k_b = 0$); C) reputational utility associated with being perceived as impartial, with no weight placed on justice concerns ($k_j = 0$); D) communicative utility of harsh punishment, mild punishment and inaction as a function of the polarization parameter p and m^* which denotes the authority's own moral judgment about the target act.

Polarization in legitimacy beliefs

In this setting, the audience consists of two groups who hold shared beliefs about the wrongness of the target act, but are polarized in their prior beliefs about the authority's legitimacy. The in-group strongly views the authority as highly motivated by justice and impartial, and the out-group initially distrusts the authority, viewing them as weakly motivated by justice and biased towards the target. We express the degree of social fracture by holding the in-group's beliefs about the authority fixed ($\mu_j = 0.9$, $\mu_b = 0.0$) and using a polarization parameter $p \in [0, 1]$ to control the out-group's level of distrust. Specifically, we linearly interpolated the out-group's priors from high trust, matching the in-group ($\mu_j = 0.9$, $\mu_b = 0.0$) at $p = 0$, to high distrust ($\mu_j = 0.1$, $\mu_b = \pm 0.4$) at $p = 1$. We assume that the out-group's beliefs about the authority's bias track the authority's relationship to the target: a rival is expected to be biased against the target, whereas an ally is expected to be biased in favor, yielding $\mu_b = -0.4$ and $\mu_b = 0.4$, respectively. The variance is fixed at approximately 0.01. We orthogonally varied the mean of the audience shared belief about wrongness (0.5-0.8), keeping the variance constant at 0.028.

Results

Polarization in moral beliefs

Reputational concern When audiences are polarized in their moral beliefs but hold shared uncertain beliefs about the authority, reputational concerns favor a moderate punitive policy: mild punishment is preferred to both harsh punishment and inaction (Fig. 1A). In fact, both dimensions of legitimacy in the model, justice and impartiality, push the reputation-aware authority toward mild punishment (Fig. 1B–C). As the two groups become more extreme in their moral beliefs, mild punishment loses its utility as a proportionate response; and, harsh punishment or inaction appear more proportional to one group yet highly unjust and biased to the other. As a consequence, reputational benefits with one group are offset by reputational costs with the other group and mild punishment remains preferable overall. However, as the utilities of the three responses converge, reputational concerns carry less effective weight in shaping the authority's punitive policy.

Notably, reputational utility is independent of the authority's own belief about wrongness (m^*), as well as of how much the authority intrinsically values proportionality or the target's welfare. As a result, reputational concerns can conflict with the authority's intrinsic motives–for instance, if the authority personally views an act as highly wrong and desires harsh punishment, yet reputational considerations dictate a milder response.

Communicative concern Under communicative goals, the preferred punitive policy depends on both the level of polarization and the authority's own moral belief, m^*. When the authority holds a moderate view ($m^* = 0.5$), mild punishment is the optimal strategy across all levels of polarization (Fig. 1D). Although the utility of all actions decline as the society becomes more polarized, mild punishment shifts both groups' moral beliefs toward the authority's own view more effectively than either harsh punishment or inaction.

For authorities with more extreme views about the wrongness of the act ($m^* = 0.7$ and $m^* = 0.9$), harsh punishment becomes more attractive, especially when audiences are relatively moderate (Fig. 1D), because it effectively commu-

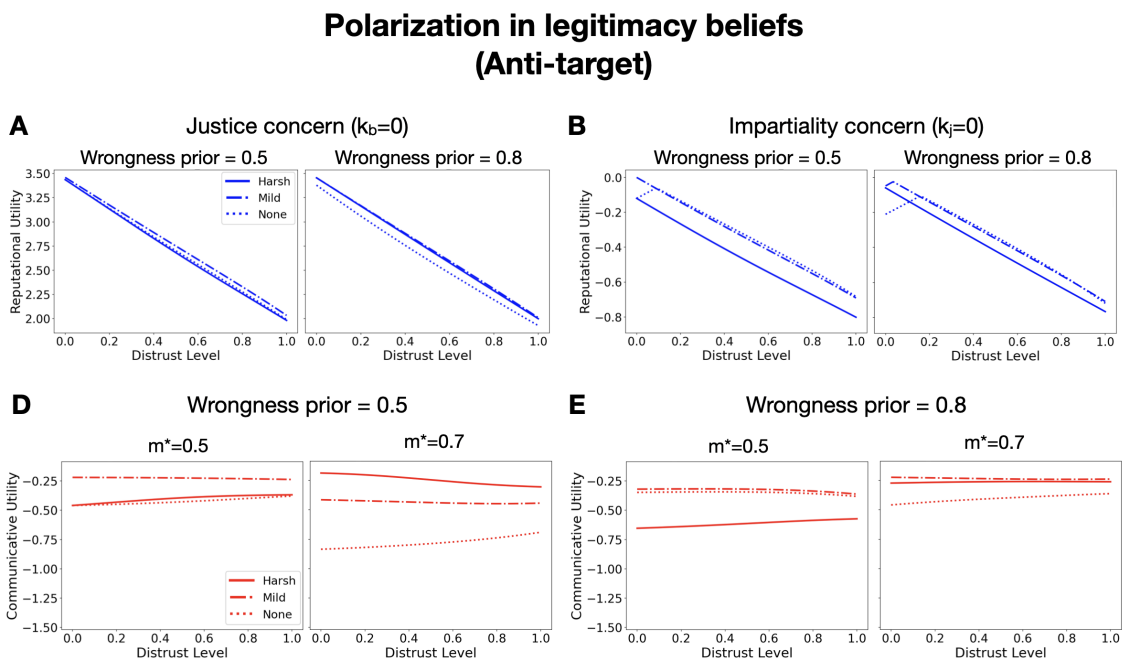

Figure 2: Reputational and communicative utility of punishment when facing audiences polarized in their beliefs about the authority. A) reputational utility associated with being perceived as just for harsh punishment, mild punishment, and inaction as a function of the polarization parameter p that controls the level of outgroup's distrust in the authority, wrongness prior is the mean of audience shared belief distribution about wrongness; B) reputational utility associated with being perceived as impartial; D) communicative utility of harsh punishment, mild punishment and inaction as a function of the polarization parameter p, audience prior belief about wrongness, and m^* which denotes the authority's own moral judgment about the target act.

nicates to both groups that the act is more wrong than they currently believe. However, as polarization of moral beliefs increases, the utility of harsh punishment declines relative to mild punishment and inaction for two reasons. First, the out-group initially views the act as relatively innocuous, so harsh punishment signals that the authority is not motivated by justice (Fig. 1B). This, in turn, reduces the informativeness of harsh punishment about wrongness, so the out-group's beliefs about wrongness shift only weakly. Second, with more moral polarization, the in-group's prior belief moves closer to m^* already, or is more extreme than the authority. In these cases, harsh punishment does not effectively communicate the authority's moral beliefs to the audience.

Polarization in legitimacy beliefs

Reputational concern When audiences are polarized in their legitimacy beliefs, reputational utility depends on the level of the outgroup's distrust in the authority, the direction of the authority's bias (against or in favor), as well as the audience's shared belief about the wrongness of the target act.

A concern to be perceived as just by the audience mainly pushes the authority toward a punitive response that is proportionate to the audience's shared belief about wrongness (Fig. 2A). For example, when the audience believe the act is only moderately wrong (i.e., mean of wrongness prior distribution=0.5), mild punishment yields the highest utility; whereas, harsh punishment becomes more attractive as the audience initially hold more extreme beliefs about the wrongness of the act (mean of wrongness prior=0.8).

On the other hand, a concern to be perceived as impartial primarily creates an incentive for authorities to be more lenient toward opponents (Fig. 2B) and harsher towards allies (not shown), especially at high levels of out-group distrust. When distrust is low, both groups view the authority as highly just and nearly impartial, so mild punishment best preserves perceptions of impartiality, whereas harsh punishment and inaction signal bias against and in favor of the target, respectively. As out-group distrust increases, inaction gains utility for an anti-target authority because it shifts the out-group's prior belief toward greater impartiality. This advantage eventually saturates, however, since inaction simultaneously makes the in-group more likely to infer pro-target bias, while at high levels of distrust it cannot fully overturn the out-group's initial suspicion of bias.

As the audience's shared belief that the act is wrong increases, two effects emerge. First, higher levels of distrust are required for inaction to become preferable to mild or harsh punishment. Second, harsh punishment gains utility across all levels of distrust, making the utilities of the three responses converge and reducing the effective weight of impartiality concerns in shaping the authority's policy. These effects reverse for a pro-target authority: harsh punishment becomes preferred to both mild punishment and inaction, and this effect strengthens as the audience's belief in the act's wrongness increases, amplifying the influence of impartiality concerns.

Communicative concern As both groups share the same prior over wrongness, the ordering of punitive responses is

determined primarily by the authority's own moral belief (m^*) relative to the audience's shared belief about the wrongness of the act. Importantly, communicative utility cannot be reduced to a preference for punitive responses proportionate to the authority's own moral belief. Rather, it depends on the interaction between the authority's moral belief and the audience's prior belief about wrongness. For example, when $m^* = 0.5$ and the audience's prior wrongness belief is high ($W = 0.8$), mild punishment and inaction can have similar communicative utility despite their different levels of proportionality (see Equation 1 for comparison with U_{justice}).

High levels of out-group distrust reduce the effective weight of communicative concerns in the authority's policy without changing the relative evaluation of the three actions. Due to reduced trust, punishment becomes less effective at communicating wrongness to the out-group, and the response utilities become differentiated primarily by their influence on the in-group.

Discussion

Authorities, such as institutional leaders, managers and teachers, often make punitive interventions while facing diverse audiences polarized in both their moral beliefs and their trust in the authority. In such contexts, the same punishment can have contrasting consequences in terms of both communicating the wrongness of the target act and maintaining or establishing trust in the authority (Amemiya et al., 2020; Gau and Brunson, 2015; L. Mulder, 2018; Radkani et al., 2025). We developed a cognitively inspired model of strategic punitive decision-making in which authorities anticipate and seek to shape audience inferences along both dimensions. Simulations of the model generated systematic predictions about how strategic goals and contextual features shape punitive decisions, and when these pressures come into conflict.

We found that the influence of communicative and reputational concerns on punitive decisions depends sensitively on the form and degree of audience polarization, as well as the authority's own moral convictions. Depending on these contextual features, communicative and reputational goals may align with one another or come into conflict, both with each other and with the authority's non-strategic motives. For example, in societies polarized in their moral beliefs, reputational concerns promote a moderate punitive policy, which may conflict with the authority's sense of justice when they personally view the act as highly wrong and prefer harsh punishment. Depending on the level of polarization, communicative goals may also align with reputational goals or support harsh punishment instead. Similarly, in societies polarized in their trust towards the authority, concerns about being perceived as just, concerns about impartiality, and communicative goals may each support different punitive policies.

Existing empirical work suggests that punitive decisions are highly dependent on the context, including who the audience is (Jordan and Kteily, 2025), what they know about the transgression (de Kwaadsteniet et al., 2019), and the authority's relationship to the target (Tsai, 2021). Our computational model synthesizes these findings by formalizing how strategic authorities incorporate such contextual features when planning punishment. It also goes beyond existing paradigms by revealing the potential complexity of punitive decision-making under audience diversity, and generating systematic predictions for these previously untested contexts. An important direction for future work is to test whether and how authorities are sensitive to these contextual features and the conflicts they generate in planning their punitive policies.

Although the model and simulations are developed in the context of punishment, they can be extended to other forms of authority intervention, such as debunking misinformation, content moderation, or public condemnations. Just as punishment signals a norm violation, fact-checks or warning labels signal epistemic violations while also conveying information about the authority's hidden motives (Reinero et al., 2023; Benegal and Scruggs, 2018). Similar trade-offs may therefore characterize authority decisions across these domains beyond punishment (Radkani et al., 2024). Future work should examine whether and how audiences infer authorities' strategic motives, for instance by tracking the systematic variation in their policies across contexts.

Model limitations and extensions We made several simplifying assumptions in our computational model and simulations. First, we represented the audience as two groups that are polarized along some dimensions while sharing others, in order to isolate their contributions to authority decision-making. Future work should consider richer settings, including multiple groups or societies that are simultaneously polarized in both moral beliefs and trust in the authority. These forms of polarization are likely to interact, since audience inferences depend jointly on beliefs about wrongness and beliefs about the authority (Radkani et al., 2025). We also assumed that the authority equally values either reputation or communication to both groups. In real world settings, authorities may have different goals for each group, for example wanting to sustain reputation with the in-group while shifting the norm beliefs of the out-group.

Second, we assumed that audiences agree on the meaning and severity of each punitive action. In practice, interpretations of the same action often vary due to differences in priors or cultural knowledge. Authorities can exploit such ambiguity through "dog whistling," selectively conveying meaning to in-groups while appearing neutral to out-groups (Albertson, 2015). A natural extension of our framework can capture this phenomenon by allowing the perceived target cost (U_{target}) to vary as a function of the audience identity.

Finally, we focused on a single intervention, whereas beliefs about norms and legitimacy are typically shaped through repeated interactions with an authority. Future work should examine how strategic authorities plan over longer horizons by anticipating the evolution of audience beliefs over time and the interactions among different belief dimensions (Radkani et al., 2024).

Acknowledgments

We thank Joshua Tenenbaum for teaching the Computational Cognitive Science course from which this project originated, and Marlene Berke for her encouragement to develop the initial course assignment into this publication. H.X.M. was supported by the Jack Kent Cooke Foundation. R.S. was supported by the Patrick J McGovern Foundation and the Guggenheim Foundation. S.R. was supported by the National Institutes of Health Training Grant in Computational Psychiatry (T32MH126388). ChatGPT was used for editing and streamlining the writing.

References

Albertson, B. L. (2015). Dog-whistle politics: Multivocal communication and religious appeals. *Political Behavior*, *37*, 3–26. https://doi.org/10.1007/s11109-013-9265-x

Amemiya, J., Mortenson, E., & Wang, M.-T. (2020). Minor infractions are not minor: School infractions for minor misconduct may increase adolescents' defiant behavior and contribute to racial disparities in school discipline. *American Psychologist*, *75*(1), 23.

Ashokkumar, A., Galaif, M., & Swann, W. B. (2019). Tribalism can corrupt: Why people denounce or protect immoral group members. *Journal of Experimental Social Psychology*, *85*, 103874. https://doi.org/https://doi.org/10.1016/j.jesp.2019.103874

Benegal, S., & Scruggs, L. (2018). Correcting misinformation about climate change: The impact of partisanship in an experimental setting. *Climatic Change*, *148*. https://doi.org/10.1007/s10584-018-2192-4

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? deterrence and just deserts as motives for punishment. *Journal of personality and social psychology*, *83*(2), 284–299. https://doi.org/10.1037/0022-3514.83.2.284

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.

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.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140. https://doi.org/10.1038/415137a

Frostig, R., Johnson, M., & Leary, C. (2018). Compiling machine learning programs via high-level tracing. https://mlsys.org/Conferences/doc/2018/146.pdf

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*(8), 986–997. https://doi.org/10.1177/0146167214533130

Gau, J., & Brunson, R. (2015). Procedural injustice, lost legitimacy, and self-help. *Journal of Contemporary Criminal Justice*, *31*, 132–150. https://doi.org/10.1177/1043986214568841

Goodman, N., & Frank, M. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*. https://doi.org/10.1016/j.tics.2016.08.005

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

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), 520–549. https://doi.org/10.1037/xge0000569

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. https://doi.org/10.1038/nature16981

Jordan, J. J., & Kteily, N. S. (2025). Punitive but discerning: Reputation can fuel ambiguously deserved punishment, but does not erode sensitivity to nuance. *Journal of personality and social psychology*, *128*(5), 1072–1102. https://doi.org/10.1037/pspa0000435

Marshall, J., Mermin-Bunnell, K., & Bloom, P. (2020). Developing judgments about peers' obligation to intervene. *Cognition*, *201*, 104215.

Mulder, L. (2018). When sanctions convey moral norms. *European Journal of Law and Economics*, *46*. https://doi.org/10.1007/s10657-016-9532-5

Mulder, L. B., Verboon, P., & Cremer, D. D. (2009). Sanctions and moral judgments: The moderating effect of sanction severity and trust in authorities. *European Journal of Social Psychology*, *39*, 255–269. https://api.semanticscholar.org/CorpusID:144764230

Nosenzo, D., Xiao, E., & Xue, N. (2024). The motive matters: Experimental evidence on the expressive function of punishment. *Games and Economic Behavior*, *148*, 44–67.

Radkani, S., Landau-Wells, M., & Saxe, R. (2024). How rational inference about authority debunking can curtail, sustain, or spread belief polarization. *PNAS Nexus*, *3*(10), pgae393. https://doi.org/10.1093/pnasnexus/pgae393

Radkani, S., Tenenbaum, J., & Saxe, R. (2022). Modeling punishment as a rational communicative social action. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*.

Radkani, S., Tenenbaum, J. B., & Saxe, R. (2025). What people learn from punishment: A cognitive model. *Proceedings of the National Academy of Sciences of the United States of America*, *122*(32), e2500730122. https://doi.org/10.1073/pnas.2500730122

Rai, T. S. (2022). Material benefits crowd out moralistic punishment. *Psychological science*, *33*(5), 789–797. https://doi.org/10.1177/09567976211054786

Raihani, N. J., & Bshary, R. (2015). The reputation of punishers. *Trends in ecology evolution*, *30*(2), 98–103. https://doi.org/10.1016/j.tree.2014.12.003

Raihani, N. J., & Bshary, R. (2019). Punishment: One tool, many uses. *Evolutionary Human Sciences*, *1*, e12.

Reinero, D., Harris, E., Rathje, S., Duke, A., & Bavel, J. (2023, May). *Partisans are more likely to entrench their beliefs in misinformation when political outgroup members fact-check claims*. https://doi.org/10.31234/osf.io/z4df3

Salcedo, J. C., & Jimenez-Leal, W. (2024). Severity and deservedness determine signalled trustworthiness in third party punishment. *British Journal of Social Psychology*, *63*(1), 453–471.

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. https://doi.org/10.1016/j.cognition.2020.104544

Saxe, R. (2022). Perceiving and pursuing legitimate power. *Trends in cognitive sciences*, *26*(12), 1062–1063. https://doi.org/10.1016/j.tics.2022.08.008

Sun, B., Jin, L., Yue, G., & Ren, Z. (2023). Is a punisher always trustworthy? in-group punishment reduces trust. *Current Psychology*, *42*(26), 22965–22975.

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.

Tyler, T. R. (2006). Psychological perspectives on legitimacy and legitimation. *Annual review of psychology*, *57*, 375–400. https://doi.org/10.1146/annurev.psych.57.102904.190038

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.

Xiao, E. (2013). Profit-seeking punishment corrupts norm obedience. *Games and Economic Behavior*, *77*(1), 321–344.