Received: 29 March 2024 | Revised: 13 July 2024 | Accepted: 17 July 2024

DOI: 10.1111/spc3.12992

REVIEW ARTICLE

WILEY

Cooperation, domination: Twin functions of third-party punishment

Jordan Wylie[1] | Ana Gantman[2,3]

[1]Boston College, Chestnut Hill, Massachusetts, USA

[2]The Graduate Center, City University of New York, New York, New York, USA

[3]Brooklyn College, New York, New York, USA

Correspondence
Jordan Wylie, Department of Psychology & Neuroscience, Boston College, Chestnut Hill, MA 02467, USA.
Email: wyliej@bc.edu

Funding information
National Science Foundation

Abstract

Rules serve many important functions in society. One such function is to codify, and make public and enforceable, a society's desired prescriptions and proscriptions. This codification means that rules come with predefined punishments administered by third parties. We argue that when we look at how third parties punish rule violations, we see that rules and their punishments often serve dual functions. They support and help to maintain cooperation as it is usually theorized, but they also facilitate the domination of marginalized others. We begin by reviewing literature on rules and third-party punishment, arguing that a great deal of punishment research has neglected to consider the unique power of codified rules. We also argue that by focusing on codified rules, it becomes clear that the enforcement of such rules via third-party punishment is often used to exert control, punish retributively, and oppress outgroup members. By challenging idealized theory of rules as facilitators of social harmony, we highlight their role in satisfying personal punishment motives, and facilitating discrimination in a way that is uniquely justifiable to those who enforce them.

KEYWORDS
morality, norms, rules, third-party punishment

1 | INTRODUCTION

Rules, including both explicitly codified rules and implicit social norms, are theorized to solve a number of coordination problems (Lewis, 1986; see also Singh et al., 2017), which, in essence, explain the flourishing of institutions and cooperation in large-scale human societies (e.g., Fehr & Schurtenberger, 2018; Henrich & Muthukrishna, 2021; Heyes, 2024). They are pervasive—present in almost all parts of our lives. They show up early in life (e.g., Grueneisen & Tomasello, 2019; Nobes, 1999; Göckeritz et al., 2014; Hardecker et al., 2017; Piaget, 1932/2013), often spontaneously; rules govern how we play games, participate in sports, interact with one another in society (i.e., the law), and more. Rules, then, facilitate and maintain complex human societies, helping humans to cooperate, organize each other, and maintain positive social order.

This is (also) what comes to mind for most people when they are asked to think about the function of rules. For example, in pilot data collected for a separate manuscript from 102 consenting Prolific participants, we asked participants to tell us what rules are for, their purpose, and whether they apply to everyone. Words like "order", "maintain", "fair", "society", "chaos", and "safe" are in the top 10 most important words used across participants (using tf-idf to analyze word importance).[1,2] People often see rules as essential tools for maintaining social order, fairness, and safety.

Yet, this is not the complete story about what the rules are for—rules, particularly official, codified rules, can be enforced (i.e., their violation comes with a punishment). These punishments can have many uses (Raihani & Bshary, 2019), and be wielded to shape life outcomes, often in punitive ways. The ensuing punishments that follow rule-breaking, often administered by third parties and institutions, can be some of the harshest punishments our societies can conjure, including death and solitary confinement (see e.g., Bernhard et al., 2022).

Here, we explore the ways that codified rules have twin functions: cooperation and the maintenance of order on the one hand, and domination and control on the other. From the point of view of history our point is trivially obvious: the bureaucratization of social living comes with tradeoffs, famously between equality and freedom (Weber, 1921), compliance and character Arendt, 1963), and work and imagination (Graeber, 2015). Official rules and their pre-specified state-sanctioned enforcement have long been used intentionally to marginalize, oppress, and systematically murder members of stigmatized groups (e.g., the Virginia Slave Codes of 1705, which codified and protected the practice of the enslavement of Africans and their children by law in the United States, the Indian Removal Act of 1830, which legalized the forced relocation and murder of Native Americans in the United States, and the Final Solution of 1941–1945, a euphemism for the Nazi's administrative plan to murder all Jews within their reach).Yet, theorizing about the functions of third-party punishment have not sufficiently been put in conversation with these historical events and empirical findings in keeping with them (which we will review below). It is the aim of this paper to review and synthesize work on the function of third-party punishment as it operates in society in order to highlight the punitive and oppressive function that official third-party punishment serves.

In this way, we highlight the dark, twin function of codified rules that comes alongside cooperation: domination. To do this, we will first define codified rules with an emphasis on their inextricable link to third party punishment, and their distinctiveness from social norms. Then we will review the importance of third-party punishment for cooperation on the one hand, and oppression on the other. We conclude that these two functions of codified rules–and the third-party punishments that their violations entail–cannot be ordered. That is, one function does not supersede the other in society, and should not, in our theorizing.

2 | WHAT ARE CODIFIED RULES

People care greatly about enforcing the rules, though some more than others (Adorno et al., 1950; Altemeyer, 1981, 1988, 1996; Gelfand et al., 2006, 2011; Whitley Jr, 1999). People are generally willing to incur a cost to enforce the rules (Buckholtz, et al., 2008; Fehr & Fischbacher, 2003; Jensen et al., 2007; Riedl et al., 2012),

find it rewarding (see Darley, 2009 for review), and believe that enforcing the rules is necessary for the maintenance of social order (Tyler & Boeckmann, 1997). That is, societal cohesion and stability heavily depend on the enforcement of rules. Rules foster cooperation and allow for complex societies and cooperation across non-kin (Fehr & Gächter, 2002; see also Fehr & Fischbacher, 2004; Buckholtz, et al., 2008; Jensen et al., 2007; Riedl et al., 2012).

While rules can be unwritten (e.g., many social norms), we focus here on codified rules because the world is awash in them (Graeber, 2015), and as scientists and as people we tend to take their positive impact, especially on large-scale societies, for granted. Here, we take explicitly codified rules to have a few defining properties:

1. They are explicit directives for how to or not to behave
2. They are enforceable by a third party:
   a. That is part of a larger institution
   b. Which pre-specifies the possible punishments
   c. And a representative of that institution implements the punishment
   d. There is a reporting mechanism by which the second party can report the rule violation to the enforcing third party
3. They are meant to apply equally to all within their jurisdiction

This kind of rule is an explicit directive that outlines how to (or not to) behave.[3] In this paper, we focus exclusively on these explicitly codified rules, like laws (proscribing identity theft, jaywalking) or the official rules of a game (white goes first in chess, the ball has to land on or within the lines to be playable in tennis) or place (no one under 21 allowed in the bar). Codified rules show up in everything from the unread terms and conditions when we download new apps on our phones to the laws proscribing how people may treat each other (e.g., rules against murder, unwanted sexual contact).

Cross-culturally, people have expectations for how laws (one kind of rule) can and should behave. For example, Hannikainen et al. (2021) looked across 11 countries and found that people generally believe laws do not and cannot contravene certain fundamental legal principles. People largely agree that the law is not and could never be secret (i.e., not public information), unstable (changing frequently), apply to behaviors that were legal when conducted, apply to behaviors which are impossible, and critically, apply only to specific people. That is, both in the abstract and in how people think the law actually behaves, laws (and other rules) are meant to apply generally.

Critically, codified rules are enforceable by a third party. When the ball falls outside the lines in tennis, a line judge (or computer program) calls it as such, and the point is over. A bouncer will block minors from entering the bar if they are underage. In other words, explicit codes of conduct function such that their violations incur pre-specified punishments dispensed by an unaffected third party (i.e., third-party punishment; Fehr & Fischbacher, 2004; Buckholtz, et al., 2008; Jensen et al., 2007; Riedl et al., 2012). These punishments can come in many forms, including in the form of loss of money (e.g., Fehr & Fischbacher, 2003), a fine or penalty (Bögelein, 2017), or a loss of one's time (e.g., required volunteer service), and are administered by some representative of an institution[4] (e.g., the U.S. legal system; Searle, 1995). The range of punishments for breaking these explicit rules is specified in advance. For example, in New York State, the current fine for jaywalking is up to $250. The penalty for hitting the ball out of bounds in tennis is the loss of a point. In this way, rules, and the institutions that codify and enforce them, are a way of moving some circumscribed activity (tennis, civic life, chess) from one that is enforced by a second party, the agent or system that is directly affected, to one that is enforced by a third party. These rules critically proscribe actions, not people, and people expect punishments to follow their violation.

3 | THE UNIQUE POWER OF CODIFIED RULES

Much of the literature on third-party punishment focuses on the enforcement of social norms. Yet, the implications of the prevalence of third-party punishment among humans is meant, at least in part, to help us understand how and why we have come to have institutions that formalize third party punishment. We contrast codified rules with social norms exactly because the appropriate means for punishing codified rules and social norms are distinct—when people break official rules, we expect their punishment to come from the institution that codified the rule in the first place. When people violate a social norm, we expect that their punishment will come in the form of social sanctioning (e.g., Wylie & Gantman, 2023). This demonstrates one of the ways that codified rules are special. Though some define social norms very broadly to include codified rules (Heyes, 2024), we see them as distinct (see also Bicchieri et al., 2023), and we separate them here because we think they are meaningfully distinct and they interact: Formal rules can be used to codify desired norms (Posner, 1997), shape norms by increasing social disapproval (Lane et al., 2023), and can also shift norms in a top-down manner (Tankard & Paluck, 2017).[5]

Much like codified rules, social norms also have a marked influence on society and everyday behavior. But these informal rules lack the properties that make codified rules especially powerful tools. Where norms are flexible, codified rules are rigid and given power by important institutions in society. Norms are often directly observable and inform behavior in a bottom-up manner; they don't require a top-down supervision (Young, 2015). Instead, norms emerge and evolve through social and cultural learning: humans are incredibly adept at using observations of what others do to inform how to behave (see Constantino et al., 2022 for review). In contrast, codified rules are invoked in a top-down manner. They are written down—specified by authoritative entities so they are easily articulated and applied where applicable. This makes codified rules seem legitimate, which can increase compliance (Gray & Roberts-Gray, 1979). Codified rules also come with predefined punishments, which reduce the costs associated with punishment or enforcement (Kube & Traxler, 2011[6]), and may even embolden people to bring up their noncompliance (Ellickson, 1991). When you notice someone doing something undesirable, like not picking up after their dog, it can be difficult to articulate to strangers why it is wrong. But when that transgression is codified, now you can simply state that it is "against the law". That is, by virtue of them being codified by a relevant authority, they reduce costs and risks—to both an affected second party and third-party observers.

Formal rules are, however, often more rigid than their informal counterparts—their alteration requires processes that can be lengthy and bureaucratic. For example, the Jones Act of 1920 still influences how cargo is transported by sea in the US today. A policy passed a century ago continues to dictate the logistics of cargo transportation, though it is likely that it is not well optimized to today's world. And yet there are many such policies and legacy laws created decades ago that may seem archaic today but still technically hold power. Norms, on the other hand, frequently evolve as society and culture changes. Their dynamic nature is part of what makes them so pervasive and informative. But this flexibility also makes them less stable (unless they become codified; see Posner, 1997) and susceptible to tipping and erosion (Bicchieri et al., 2022; Keizer et al., 2008; Legros & Cislaghi, 2020). This means that many informal rules and norms from a century ago have lost their relevance today, certain laws from the same era still impact our lives—underscoring the longevity and resilience of formal rules.

Codified rules uniquely formalize the separation not only between the person who receives the punishment and the harmed party (if there is one) but also, in some cases, between a third-party observer of a transgression and its punishment. Take for example, a person who decides to call 311 (a municipal service expressly for non-emergency complaints) in New York City after seeing someone graffiti in an alleyway. In this case, the person who broke the rule did so in front of a third person, not directly harmed by their graffiti, and then the NYPD arrives to determine what should happen next. Maybe the person gets a fine or in rare cases even ends up arrested (Wylie et al., 2024). The codification of the rule against graffiti separates the person spray painting, from the person who called 311, from the punishment.

This is critically different from norm enforcement in which the person who wants to enforce the norm often must themselves engage in interpersonal sanctioning, gossip, ostracism, or some other way of communicating that

something wrong was done. Even indirect punishment like gossip, entails higher stakes for the enforcer than picking up the phone to dial 311. People become deputized as actors of the institutions that codify and enforce the rules (Yankah, 2024). Once deputized, we see a breakdown of the purported functions of punishment and rules in how they are actually enforced: first, people have a clear rationale to others and to themselves about why the punishment must occur (someone broke a rule after all); justification comes easily. Second, people do not have to take on the full magnitude of enforcing a rule themselves no matter how aversive the punishment (Bernhard et al., 2022). Third, people need not take on the full weight of the communication that comes with the punishment (e.g., Cushman et al., 2019). As we have formalized (i.e., by forming institutions) the role of rules, we have removed enforcers from some of the purported functions of both rules and their punishments.

This lens also allows us to notice and study the times when informal rules or norms and codified rules conflict (see focus theory of normative conduct; Cialdini et al., 1991). Take, for example, the case of phantom rules. Phantom rules represent a subclass of codified rules in which a proscribed behavior and the descriptive norm associated with that behavior are in conflict. And while phantom rules are frequently broken and seen of little moral consequence (Wylie & Gantman, 2023), social norm information is salient, and it influences individuals' behavior (e.g., Asch, 1956; Miller & Prentice, 1996; Schultz et al., 2007). On a narrow definition of social norm violations, they are not technically against any particular codified rule, but they tend to be costly interpersonally and can lead to indirect punishments such as gossip or ostracism (e.g., Molho et al., 2020). Yet, when faced with both a broken codified rule and a broken social norm, research suggests that people frequently choose to punish the codified one (in lieu of another option; Wylie & Gantman, 2023). That is, while the social norm violations may motivate the desire to punish and correct behavior, having a quick and easy tool to implement to satisfy that desire is often preferable.

Both rules and norms serve to regulate human behavior. But codified rules are special: They intersect with powerful institutions in ways that cognitive and social scientists should seek to explore and better understand (see also Saxe, 2022). They also provide another rich avenue to explore the motivations and consequences of third-party punishment.

In sum, official, codified rules, including those prescribing violence, derive their power from our readiness to codify, follow, and enforce them. They can delineate group boundaries and provide an outlet for punishment motivations that might otherwise be difficult or costly to satisfy. And they reveal the dual nature of rules as instruments for social regulation and mechanisms for exerting control.

4 | WHAT IS PUNISHMENT FOR?

Research spanning disciplines and decades has documented the emergence, motivators, and consequences of third-party punishment—punishment administered by an unaffected third-party (e.g., Buckholtz et al., 2008; Buckholtz & Marois, 2012; Crockett et al., 2014; Fehr & Fischbacher, 2004; Jordan et al., 2016; Krueger & Hoffman, 2016; Marshall & McAuliffe, 2022; McAuliffe et al., 2015; Raihani et al., 2010; Raihani et al., 2012; Tan & Xiao, 2018). Much of the foundational research on third-party punishment stems from efforts to understand altruistic punishment or costly punishment (e.g., Fehr & Gächter, 2000, 2002; Fowler, 2005; Henrich et al., 2006; Ostrom et al., 1992), and ultimately, cooperation (Boyd et al., 2010; Fehr & Gächter, 2002). This work suggests that costly punishment often emerges (Fehr & Gächter, 2002; Fehr & Fischbacher, 2003) and reduces free-riders and defectors (Fehr & Gächter, 2002; Fowler, 2005), especially when participation is voluntary (Hauert et al., 2007).

The presence of third-party punishment provides an explanation to the question of *why* people would cooperate when there is no reputational gain, no apparent reciprocity and no kinship (Fehr & Gächter, 2002). Having the opportunity to punish reduces free-riding (Fehr & Gächter, 2000), especially in the long-term (Frey & Rusch, 2012). Further, when multiple agents decide to punish a free-rider (or punishment is coordinated; see Molleman et al., 2019), cooperation and group payoffs often increase (Boyd et al., 2010). Research also suggests that people from groups with the ability to punish outperform those from groups lacking such mechanisms (Sääksvuori

et al., 2011). This is also true of individual payoffs: Punishment over the long-term increases payoff for individuals and for groups of individuals (Gächter et al., 2008). Meta-analytic evidence also suggests that the effects of punishment on cooperation are robust across designs (Balliet et al., 2011), and cross-cultural work suggest that this link emerges across different societies as well (Fitouchi & Singh, 2023; Henrich et al., 2006; Mathew & Boyd, 2011). This work has been so central that costly punishment, at one point, was even said to have overshadowed the original aim to understand and characterize cooperation (Colman, 2006). Nonetheless, evidence suggests that costly punishment, by both second and third parties, plays a critical role in promoting cooperation.

Alongside these advances in theory on costly punishment and cooperation (see Henrich & Muthukrishna, 2021 for review of cooperation), there was a wave of studies that questioned the primacy of costly punishment in maintaining cooperation. This included work examining the role of rewards (e.g., Balliet et al., 2011; Rand et al., 2009), reputation (Wu et al., 2016), and social networks (Rand et al., 2011) in promoting cooperation, the puzzle of antisocial punishment (or the punishment of prosocial agents Rand & Nowak, 2011; Herrmann et al., 2008; Gächter & Herrmann, 2009; Herrmann et al., 2008; Pfattheicher et al., 2014; Pfattheicher & Schindler, 2015), evidence of variation across cultures (Wu et al., 2009; see also Molho et al., 2024 for review of cross-cultural variation in norm enforcement), and the observation that engaging in costly punishment can actually result in *diminished* payoffs in cooperative games (Dreber et al., 2008). Overall, this work suggests that costly punishment is just one of many mechanisms that are capable of sustaining both cooperation and other costly behaviors (Henrich & Muthukrishna, 2021), including those that are ultimately selfish. Thus, while punishment can and does deter undesirable behavior and promote cooperation under certain circumstances, this role of punishment is just one of the functions it serves. Punishment may not always be optimal for cooperation or collective welfare.

What then is the function of punishment? Research has suggested that punishment tends to serve an expressive or communicative (Cushman et al., 2019; Duff, 2022[7]) function. That is, people sometimes use direct punishment to communicate to offenders that behavior change is needed, often at a cost to the punisher (see Molho & Wu, 2021 for review). Though direct punishment also serves to satisfy retributive motives (e.g., Carlsmith et al., 2002). Indirect punishments, on the other hand, typically act on the reputation of and information in the community about an offender (e.g., Dores Cruz et al., 2021; Feinberg et al., 2014; Molho & Wu, 2021), and those who engage in third-party punishment can reap reputational benefits (Jordan et al., 2016).

Third-party punishment (e.g., Buckholtz et al., 2008; Buckholtz & Marois, 2012; Crockett et al., 2014; Fehr & Fischbacher, 2004; Krueger & Hoffman, 2016; Marshall & McAuliffe, 2022; McAuliffe et al., 2015; Raihani et al., 2010; Tan & Xiao, 2018), is a central force behind human cooperation (Fehr & Fischbacher, 2004; but see Baumard, 2010), especially in large, WEIRD societies (see Henrich & Muthukrishna, 2021). This kind of punishment addresses some of the limitations of costly second-party punishment by extending the range of enforceable social norms; third-party sanctions allow for the regulation of behaviors that do not directly impact others (e.g., Fehr & Fischbacher, 2004). Third-party punishment also provides distance between the transgressor and punisher—this reduces risks and costs to injured second parties and serves to head off an unending cycle of punishment and revenge between two aggrieved parties (see Raihani & Bshary, 2019 for review of retaliation following punishment), but retains the ability to communicate or express the intent of direct punishment. And third party punishment can be recursive: The bouncer kicks out the underage would-be bar hopper because the bar must follow state laws prohibiting serving alcohol to minors. It is no surprise, then, that rules, embedded in larger systems of rules, comprise institutions, which codify rules and pre-specify their modes of enforcement.

5 | THE ROLE OF INSTITUTIONS AND RULES IN PUNISHMENT

Third-party punishment reveals the importance of the creation and maintenance of centralized sanctioning institutions to punishment in modern societies. Sanctioning institutions are tools which have been created to support cooperation (Lie-Panis et al., 2023), and the formal rules that guide behavior in everyday life have a sanctioning

institution built in. Institutions serve as a bridge between individual behavior and group goals and norms: Research suggests that people tend to favor the possibility of sanctioning through an institution over the absence of any sanctioning mechanism (Gürerk et al., 2006), and sanctioning institutions which are seen as legitimate support cooperation (Baldassarri & Grossman, 2011; see also Raihani & Bshary, 2019). People also tend to vote to sanction low contributors (vs. no sanction or sanctioning of high contributors; Ertan et al., 2009)—there is a preference for and stability in the existence of a fair institution for sanctioning (Kosfeld et al., 2009; Van Bavel et al., 2022).

But, of course, sanctioning institutions do not merely publish a list of possible punishments. They provide a list of rules and policies and regulations. And if those rules and policies and regulations are not met, then the punishments become available. And so rules too must play a pivotal role in large-scale human cooperation, and one that has been relatively overlooked from a psychological lens compared to the punishments they legitimize and enable. The third parties that dole out punishments—whether through costly acts by individual agents or institutions which impose sanctions—are possible because we create formal rules (see Henrich & Muthukrishna, 2021) and assume that is what those rules[8] are really for. Rules afford the formation of official sanctions and the governing institutions which uphold them—allowing for cooperation across diverse groups (De Dreu, et al., 2023).

Explicitly codified rules may become increasingly present as group sizes increase and governing institutions to enforce them emerge. Indeed, small, hunter-gatherer societies that do not have states also don't tend to have these kinds of codified rules (or very much punishment at all; Baumard, 2010). Moving from a more ultimate perspective of rules as instruments of cooperation to their more proximate functions as codes of conduct: rules are also explicit directives which are designed to proscribe actions uniformly across individuals, irrespective of their status (Weber, 1921; Graeber, 2015; Hannikainen et al., 2021; see also Wylie & Gantman, 2023). This is an important component of procedural justice, which is the perception that the rules or processes by which those rules are enforced are done so fairly (Tyler, 1997, 2003; Tyler & Lind, 1992). Thus, codified rules, by promising uniformity in enforcement and fairness in their creation, bolster the legitimacy (and authority) of these institutions (Tyler, 1994; but see Ruder & Woods, 2020). This relationship between the rules and the rule sanctioning institutions is a critical one: The effectiveness of rules in shaping behavior and norms relies on the legitimacy and sanctioning power of the institutions that enforce them (see Saxe, 2022 for a call for more work on legitimacy).

This work together highlights part of what makes codified rules powerful, and the importance of studying them within the hierarchical structures in which they emerge. They are powerful because their clear guidelines make it easy for third parties to enforce them. When third parties handle enforcement of codified rules, the justification for punishment preexists. The individual punishment for the individual rule violation does not need its own justification. In short, codified rules are essential for influencing behavior and keeping society organized. Their clear standards and enforceability are key to their role in ensuring fairness.

6 | EVEN THIRD-PARTY PUNISHMENT IS PERSONAL

As we have reviewed, the special status of rules means that they are expected to provide a universal standard of behavior regardless of individual or group affiliations. However, rules are often designed and enforced in ways that benefit those in power (Singh et al., 2017). This introduces a bias that can and does undermine their purported universality. Looking at extant research on third-party punishment, we again see that third-party punishers tend to punish in ways that align with their own interests (Krasnow et al., 2016; Rand & Nowak, 2011) or those of their ingroup. For example, research suggests that intergroup dynamics play a critical role in who people decide who to punish (e.g., Bernhard, Fehr et al., 2006; Bernhard, Fischbacher et al., 2006) and cooperate with (see Van Bavel et al., 2022). People prefer to punish norm violators when the victim is in the ingroup (Bernhard, Fehr et al., 2006) or the perpetrator is in the outgroup (Bernhard, Fischbacher et al., 2006). Meta-analytic evidence also suggests the sanctioning and enforcement of rules frequently benefits ingroup members (see Balliet et al., 2014). That is, punishment, much like parochial altruism, can serve to benefit some while harming others (Choi & Bowles, 2007),

and is sometimes even used to sanction prosocial behaviors (Herrmann et al., 2008). Everything from physical attractiveness (Li & Zhou, 2014) to group membership (Bernhard et al., 2006; Schiller et al., 2014; Yudkin et al., 2016) and complexity of the population (Marlowe et al., 2008) influences to whom and how much punishment is assigned by third-parties. Ultimately, these tendencies highlight the motivated and often parochial nature of rule enforcement, as it often aims to protect and benefit ingroup members, demonstrating a preference for maintaining social harmony but primarily *within* one's own community or group. Though these patterns are on the one hand, in keeping with a vast literature on intergroup bias (e.g., Chae et al., 2022; Kubota et al., 2013; Schiller et al., 2014; Yudkin et al., 2016), they are perhaps particularly notable here because official, codified rules are specifically meant to protect people from exactly this kind of partial evaluation (Graeber, 2015; Hannikainen et al., 2021).

Despite steady evidence for partiality in third-party punishment, its use for power maintenance and state violence has been relatively less explored from a psychological lens, especially with regard to codified rule enforcement. That is, there are obvious cases in which, for example, the law is enforced, not to promote cooperation, but to oppress and intimidate minority groups. As just one example, there is rampant racial bias in policing in the United States (Desilver et al., 2020), running as a throughline from the most severe punishments to the most mundane infractions; Perceived stereotypically of Black defendants predicts death penalty sentences (Eberhardt et al., 2006) and neighborhood racial composition predicts arrest rates for minor infractions like loud music—infractions that prima facie do not warrant arrest (Wylie et al., 2024).

Other work has found that people enforce codified rules according to their own desires and sense of who they want to see punished (Wylie et al., 2024; Wylie & Gantman, 2023), and others have found discrimination in enforcement by ethnicity and nationality (Spadaro et al., 2023) which undermines cooperation (Molenmaker et al., 2023). Further, research on aversive punishment suggests that people find punishments which minimize the harm done to the self more permissible than those which minimize the suffering of others (Bernhard et al., 2022). These findings, alongside many historical examples of abuses of state power, suggest that rules, and their enforcement, are often wielded as tools of oppression.

Further, by focusing on codified rules we can better specify our theory on the functions, both ultimate and proximate, of punishment. One such theory, punishment as communication (Cushman et al., 2019; Sarin et al., 2021), argues that a key aspect of punishment is the inferred communicative message of the punisher. That is, alongside incentives directly related to punishment, people are sensitive to what message punishers are trying to convey. When we apply this theory to the enforcement of codified rules in present society like the law, it is difficult to determine what any individual punishment communicates because the justification for the punishment comes from the whole system itself rather than the individual enforcer (and perhaps this partly underlies the need to give blank bullets to some members of the state's firing squad). However, some argue that extant modes of punishment as practiced (e.g., penal system in the United States) cannot tell us about the ultimate function of punishments because practical applications will always deviate from normative theories of how it should work (see Duff, 2022). But, of course, research on third-party punishments seeks to do exactly this: to study how punishment actually works and infer what it is for. Despite evidence that third-party punishment is subject to abuse of power, self- and group-level interests, the primary understanding of the function of third-party, and institutional punishment remains cooperation. This may be indicative that current theorizing about rules and punishment as tools for cooperation is an ideal theory—one that takes the ideal conceptualization of the idea as the model for what it actually is—leaving out practical and often unjust facts about the world (Mills, 2005).

Third-party punishment, then, extends beyond mere regulation and cooperation. It can be wielded to compete, to dominate, to favor, and to chastise. Indeed, people are more likely to choose to confront others directly and interpersonally when they have power (Molho et al., 2020). We highlight here the way that rule enforcement is subject to group-based discrimination (e.g., Schiller et al., 2014; Yang et al., 2023), the idiosyncratic punishment motives of the enforcer (e.g., Wylie & Gantman, 2023), and provide a way to assert dominance and power while upholding ideals of fairness (e.g., Pinsof, 2023). Further, people show a preference for institutional punishments for those less close to them or emotionally distant (Weidman et al., 2020), and tend to enforce codified rules beyond

their intended scope (Wylie & Gantman, 2023). Research also indicates that the willingness to engage in costly punishment may be more about exerting coercion than promoting cooperation (Dreber et al., 2008), with individuals preferring punitive actions that also enhance their reputation over those that solely improve reputation (Rockenbach & Milinski, 2006).

Taken together, we can see that punishment has many forms and serves many functions. We highlight here, a function that tends to be overlooked, especially within the context of third-party punishment. That is, retribution, domination, and hierarchy maintenance. Third parties often are in a position to enforce official codified rules, and often from a position of power. Sometimes, these punishments are administered in a discriminatory manner to serve one's group, or to serve an individual's own retributive desires. Third-party punishment, then, can be seen as a tool for maintaining social order, reinforcing existing social hierarchy, and singling out individuals for punishment.

7 | TWIN FUNCTIONS OF THIRD-PARTY PUNISHMENT

Cognitive science needs to understand, not only how rules are learned and punished in the proximate, but how rules achieve more ultimate goals of domination, power, and discrimination. These advances will serve to advance our understanding of topic areas like prejudice and stereotyping, power, and causal reasoning about structural inequality (Amemiya et al., 2023). This will also allow us to rectify disparate findings about the evolutionary origins of punishment (e.g., Dreber et al., 2008), and help us to better grasp seeming paradoxes within the punishment literature. For instance, research suggests that we have a natural aversion to inflict direct harm onto others (Crockett et al., 2014; Cushman et al., 2012) and aversion to administering punishments which we find unpleasant (Bernhard et al., 2022). Yet, we still desire to punish others and see rule-breakers get their due punishment and just deserts (Carlsmith et al., 2002). One way that humans resolve these competing motives[9] is through codification and sanctioning. Rules, then, become mechanisms through which we channel these punishment motives via third-parties. By better understanding rules, which are tools that replicate with complete fidelity as they get passed down generationally, we can better understand how systems and powerful groups across time have wielded them to carve out societies which benefit their status, access to resources, and more.

When we consider these rules in this way, it becomes clear that rules have twin functions in cooperation and domination, two forms of social coordination. Cooperation implies working toward shared goals with mutual benefit (e.g., Tomasello, 2009), yet the reality is that institutions that enforce codified rules sometimes do so to the benefit of those in power within them, and at the expense of those who are not. Thus, it is perhaps more accurate to argue that the primary function of rules is not to foster cooperation but to facilitate coordination among individuals or groups, which involves navigating power and intergroup dynamics and ensuring the smooth functioning of the institution as a whole. Groups may "cooperate" to harm, oppress, or otherwise thwart other groups, but there is something strange about calling that cooperation.[10]

8 | FUTURE DIRECTIONS

We see many fruitful avenues for future research on the function of rules as tools for domination, and retributive punishment. For instance, people may be more likely to add rules that are ambiguously enforced (e.g., phantom rules) in contexts where they already have power versus more equally enforced rules, or when they have relatively less power, as individuals or as group members. We might also predict that people defend and justify rules that are enforced in a discriminatory fashion, only when those rules benefit themselves or their group.

We have also suggested that codified rules allow people to satisfy punishment motives without incurring a social cost comparable to more direct forms of punishment (like confrontation). Accordingly, future research could

test whether people prefer to enforce official rules over opting to directly confront others after minor violations, and whether they prefer to do so anonymously.

It is institutions that codify rules and enforce them, and these institutions are created by, shape, and co-evolve with the societies that put them in place (see e.g., Henrich & Muthukrishna, 2021). Future research should explore how culture influences judgments about rule-following, rule-breaking, and rule-enforcement. For example, future work could explore how the tightness/looseness of a culture (Gelfand, et al., 2011) affects the enforcement of codified rules, in particular. Moreover, it may be that as members of a society observe discriminatory rule enforcement, this may create or justify negative stereotypes about members of groups differentially impacted by that enforcement.

There are also relevant individual differences that may influence the desire to use codified rules to oppress others and channel personal punishment motivations. For example, prior work has found that people who are more likely to endorse the view that our society needs law and order at all costs (i.e., those high in right-wing authoritarianism; Adorno et al., 1950; Altemeyer, 1981) think the police should enforce all rules, regardless of how ambiguous it is to enforce them (Wylie & Gantman, 2023; Experiment 3) and even for social norms, which are *not* codified by law (Wylie & Gantman, 2023, Supplemental Experiment S1). Prior work has also found that people with the belief that existing social hierarchy is good and should be maintained (i.e., those high in social dominance orientation; Pratto et al., 1994), are more likely to say that the police should intervene when a Black (vs. white) American breaks a rarely enforced rule like illegally parking their car (Wylie et al., 2024). Future work could examine further individual differences or investigate how these individual differences relate to the creation, maintenance, and trajectory of rule-enforcing institutions. Finally, future work could examine what factors, at the individual, situational, and societal levels lead people to create rules that *do not* benefit them directly or reduce their own individual or group-based power, and to better understand mutually beneficial cooperation that occurs within existing hierarchies.

9 | CONCLUSION

In sum, we argue that our understanding of the function of rules and their enforcement by third parties must include domination alongside cooperation. Theorizing about the ideal functions of punishment and rules has so far obscured the very real ways that third-party punishment is not for cooperation alone. Enforcement of codified rules functions to idiosyncratically punish, oppress, and dominate others. Official codified rules, and the third-party punishments their violations entail, facilitate twin functions for social coordination: both cooperation and domination.

ACKNOWLEDGMENTS

We thank Psyphi Lab and the 2023 Open Minds conference in Florence for their feedback on the early conception of this idea. We also thank our reviewers for their helpful feedback. This work was funded by an NSF (NSF SMA-2313957) awarded to the first author.

CONFLICT OF INTEREST STATEMENT

None of the authors have any conflicts of interest or financial stake in this work.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

ORCID

Jordan Wylie https://orcid.org/0000-0001-5142-8515

ENDNOTES

[1] Because of ties, there are more than 10 total words in the top ten (there are 12), the other words are: "people", "thing", "set", "general(ly)", "follow", and "world".

[2] Another way to assess oft-occurring associations with the meaning and function of rules, is to query a large language model and see what it returns. When asked "Please explain what rules are for in society" at the time of writing (March 2024, and again at revision, July 2024), both ChatGPT 4 (ChatGPT 4o in July) and Claude 3 (Claude 3.5 in July) mention "maintaining order and facilitating cooperation", "safety", and "stability". While these models are of course in flux and the purpose here is not to assess the full range of possible explanations for the existence of rules by LLMs, these responses nonetheless highlight that there is some (perhaps unsurprising) consistency, at least at the time of writing, between what people say and the outputs of models trained on human responses.

[3] Formal rules can also take other forms, such as algorithms and paradigms (see Daston, 2022), but those rules are outside of the purview of the present analysis.

[4] Here, we use institutions to refer to social systems and structures that organize society.

[5] We are carving out a conceptually narrow space, but a practically large one; codified rules control a lot of people's lives and can be extremely consequential to break (e.g., if you get arrested).

[6] Of course, norms are also punished, often through indirect means like gossip (Molho et al., 2020).

[7] While normative theory on punishment does not focus on the way that punishment works in, for example, the penal system in the United States, cognitive and social scientists *should* seek to characterize and theorize about how punishment works in various real-world contexts, including attempting to make sense of the way that punishment currently does or does not express or communicate disapproval in the penal system.

[8] Notably, rules are more universal than punishment (or third-party punishment).

[9] Not to mention the desire to be protected from second-party punishment.

[10] It is possible that the paradox of using cooperation to describe actions like oppression may go away if we consider a more pluralistic approach to morality (see also Henrich & Muthukrishna, 2021). In non-WEIRD contexts where morality is more geared toward the group rather than about harm, "cooperation-to-harm" behaviors may not represent a paradox at all. Thus, by considering not only the history and current practice of punishment (as we have done here), but also by considering the variation in moral values across people and time, we can better understand the ultimate functions of punishment as it is in reality, not in theory.

REFERENCES

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., Sanford, N., & Aron, B. (1950). *The authoritarian personality*. Harper & Row.

Altemeyer, B. (1996). *The authoritarian Specter*. Harvard University Press.

Altemeyer, R. A. (1981). *Right-wing authoritarianism*. Univ. of Manitoba Pr.

Altemeyer, R. A. (1988). *Enemies of freedom: Understanding right-wing authoritarianism*. Jossey-Bass.

Amemiya, J., Mortenson, E., Heyman, G. D., & Walker, C. M. (2023). Thinking structurally: A cognitive framework for understanding how people attribute inequality to structural causes. *Perspectives on Psychological Science*, *18*(2), 259–274. https://doi.org/10.1177/17456916221093593

Arendt, H. (1963). *Eichmann in Jerusalem: A report on the banality of evil*. Viking Press.

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and applied*, *70*(9), 1–70. https://doi.org/10.1037/h0093718

Baldassarri, D., & Grossman, G. (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences*, *108*(27), 11023–11027. https://doi.org/10.1073/pnas.1105456108

Balliet, D., Mulder, L. B., & Van Lange, P. A. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, *137*(4), 594–615. https://doi.org/10.1037/a0023489

Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin*, *140*(6), 1556–1581. https://doi.org/10.1037/a0037737

Baumard, N. (2010). Has punishment played a role in the evolution of cooperation? A critical review. *Mind & Society*, *9*(2), 171–192. https://doi.org/10.1007/s11299-010-0079-9

Bernhard, H., Fehr, E., & Fischbacher, U. (2006a). Group affiliation and altruistic norm enforcement. *The American Economic Review*, *96*(2), 217–221. https://doi.org/10.1257/000282806777212594

Bernhard, H., Fischbacher, U., & Fehr, E. (2006b). Parochial altruism in humans. *Nature*, *442*(7105), 912–915. https://doi.org/10.1038/nature04981

Bernhard, R. M., Cushman, F., & LeBaron, H. (2022). The paradox of aversive punishment. https://doi.org/10.31234/osf.io/tcsve

Bicchieri, C., Dimant, E., Gächter, S., & Nosenzo, D. (2022). Social proximity and the erosion of norm compliance. *Games and Economic Behavior*, *132*, 59–72. https://doi.org/10.1016/j.geb.2021.11.012

Bicchieri, C., Muldoon, R., & Sontuoso, A. (2023). Social norms. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/entries/social-norms/

Bögelein, N. (2017). 'money rules': Exploring offenders' perceptions of the fine as punishment. *British Journal of Criminology*, *58*(4), 805–823. https://doi.org/10.1093/bjc/azx044

Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*, *328*(5978), 617–620. https://doi.org/10.1126/science.1183665

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron*, *60*(5), 930–940. https://doi.org/10.1016/j.neuron.2008.10.016

Buckholtz, J. W., & Marois, R. (2012). The roots of Modern Justice: Cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, *15*(5), 655–661. https://doi.org/10.1038/nn.3087

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

Chae, J., Kim, K., Kim, Y., Lim, G., Kim, D., & Kim, H. (2022). Ingroup favoritism overrides fairness when resources are limited. *Scientific Reports*, *12*(1), 4560. https://doi.org/10.1038/s41598-022-08460-1

Choi, J. K., & Bowles, S. (2007). The coevolution of parochial altruism and war. *Science*, *318*(5850), 636–640. https://doi.org/10.1126/science.1144237

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. *Advances in Experimental Social Psychology*, *24*, 201–234. Academic Press. https://doi.org/10.1016/s0065-2601(08)60330-5

Colman, A. M. (2006). The puzzle of cooperation. *Nature*, *440*, 744–745. https://doi.org/10.1038/440744b

Constantino, S. M., Sparkman, G., Kraft-Todd, G. T., Bicchieri, C., Centola, D., Shell-Duncan, B., Vogt, S., & Weber, E. U. (2022). Scaling up change: A critical review and practical guide to harnessing social norms for climate action. *Psychological Science in the Public Interest*, *23*(2), 50–97. https://doi.org/10.1177/15291006221105279

Crockett, M. J., Özdemir, Y., & Fehr, E. (2014). The value of vengeance and the demand for deterrence. *Journal of Experimental Psychology: General*, *143*(6), 2279–2286. https://doi.org/10.1037/xge0000018

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion*, *12*(1), 2–7. https://doi.org/10.1037/a0025071

Cushman, F., Sarin, A., & Ho, M. (2019). Punishment as communication. In *The Oxford handbook of moral psychology* (pp. 197–209).

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. https://doi.org/10.1146/annurev.lawsocsci.4.110707.172335

Daston, L. (2022). *Rules: A short history of what we live by*. Princeton University Press.

De Dreu, C. K., Gross, J., & Romano, A. (2023). Group formation and the evolution of human social organization. *Perspectives on Psychological Science*, 17456916231179156.

DeSilver, D., Lipka, M., & Fahmy, D. (2020). 10 things we know about race and policing in the US.

Dores Cruz, T. D., Thielmann, I., Columbus, S., Molho, C., Wu, J., Righetti, F., de Vries, R. E., Koutsoumpis, A., van Lange, P. A. M., Beersma, B., & Balliet, D. (2021). Gossip and reputation in everyday life. *Philosophical Transactions of the Royal Society B*, *376*(1838), 20200301. https://doi.org/10.1098/rstb.2020.0301

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature*, *452*(7185), 348–351. https://doi.org/10.1038/nature06723

Duff, R. A. (2022). Punishment as communication.

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital-sentencing outcomes. *Psychological science*, *17*(5), 383–386. https://doi.org/10.1111/j.1467-9280.2006.01716.x

Ellickson, R. C. (1991). *Order without law: How neighbors settle disputes*. Harvard University Press.

Ertan, A., Page, T., & Putterman, L. (2009). Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review*, *53*(5), 495–511. https://doi.org/10.1016/j.euroecorev.2008.09.007

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785–791. https://doi.org/10.1038/nature02043

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87. https://doi.org/10.1016/s1090-5138(04)00005-4

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, *90*(4), 980–994. https://doi.org/10.1257/aer.90.4.980

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

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, *2*(7), 458–468. https://doi.org/10.1038/s41562-018-0385-5

Feinberg, M., Willer, R., & Schultz, M. (2014). Gossip and ostracism promote cooperation in groups. *Psychological science*, *25*(3), 656–664. https://doi.org/10.1177/0956797613510184

Fitouchi, L., & Singh, M. (2023). Punitive justice serves to restore reciprocal cooperation in three small-scale societies. *Evolution and Human Behavior*, *44*(5), 502–514. https://doi.org/10.1016/j.evolhumbehav.2023.03.001

Fowler, J. H. (2005). Altruistic punishment and the origin of Cooperation. *Proceedings of the National Academy of Sciences*, *102*(19), 7047–7049. https://doi.org/10.1073/pnas.0500938102

Frey, U. J., & Rusch, H. (2012). An evolutionary perspective on the long-term efficiency of costly punishment. *Biology and Philosophy*, *27*(6), 811–831. https://doi.org/10.1007/s10539-012-9327-1

Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: Previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1518), 791–806. https://doi.org/10.1098/rstb.2008.0275

Gächter, S., Renner, E., & Sefton, M. (2008). The long-run benefits of punishment. *Science*, *322*(5907), 1510. https://doi.org/10.1126/science.1164744

Gelfand, M. J., Nishii, L. H., & Raver, J. L. (2006). On the nature and importance of cultural tightness-looseness. *Journal of Applied Psychology*, *91*(6), 1225–1244. https://doi.org/10.1037/0021-9010.91.6.1225

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., Almaliach, A., Ang, S., Arnadottir, J., Aycan, Z., Boehnke, K., Boski, P., Cabecinhas, R., Chan, D., Chhokar, J., D'Amato, A., Subirats Ferrer, M., Fischlmayr, I. C., …, & Yamaguchi, S. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, *332*(6033), 1100–1104. https://doi.org/10.1126/science.1197754

Göckeritz, S., Schmidt, M. F. H., & Tomasello, M. (2014). Young Children's creation and transmission of social norms. *Cognitive Development*, *30*, 81–95. https://doi.org/10.1016/j.cogdev.2014.01.003

Graeber, D. (2015). *The utopia of rules: On technology, stupidity, and the secret joys of bureaucracy*. Melville House.

Gray, T., & Roberts-Gray, C. (1979). Structuring bureaucratic rules to enhance compliance. *Psychological Reports*, *45*(2), 579–589. https://doi.org/10.2466/pr0.1979.45.2.579

Grueneisen, S., & Tomasello, M. (2019). Children use rules to coordinate in a social dilemma. *Journal of Experimental Child Psychology*, *179*, 362–374. https://doi.org/10.1016/j.jecp.2018.11.001

Gurerk, O., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, *312*(5770), 108–111. https://doi.org/10.1126/science.1123633

Hannikainen, I. R., Tobia, K. P., de Almeida, G. da, Donelson, R., Dranseika, V., Kneer, M., Strohmaier, N., Bystranowski, P., Dolinina, K., Janik, B., Keo, S., Lauraitytė, E., Liefgreen, A., Próchnicki, M., Rosas, A., & Struchiner, N. (2021). Are there cross-cultural legal principles? Modal reasoning uncovers procedural constraints on law. *Cognitive Science*, *45*(8). https://doi.org/10.1111/cogs.13024

Hardecker, S., Schmidt, M. F., & Tomasello, M. (2017). Children's developing understanding of the conventionality of rules. *Journal of Cognition and Development*, *18*(2), 163–188. https://doi.org/10.1080/15248372.2016.1255624

Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A., & Sigmund, K. (2007). Via freedom to coercion: The emergence of costly punishment. *Science*, *316*(5833), 1905–1907. https://doi.org/10.1126/science.1141588

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2006). Costly punishment across human societies. *Science*, *312*(5781), 1767–1770. https://doi.org/10.1126/science.1127333

Henrich, J., & Muthukrishna, M. (2021). The origins and psychology of human cooperation. *Annual Review of Psychology*, *72*(1), 207–240. https://doi.org/10.1146/annurev-psych-081920-042106

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362–1367. https://doi.org/10.1126/science.1153808

Heyes, C. (2024). Rethinking norm psychology. https://doi.org/10.1177/17456916221112075

Jensen, K., Call, J., & Tomasello, M. (2007). Chimpanzees are rational maximizers in an ultimatum game. *Science*, *318*(5847), 107–109. https://doi.org/10.1126/science.1145850

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

Keizer, K., Lindenberg, S., & Steg, L. (2008). The spreading of disorder. *Science*, *322*(5908), 1681–1685. https://doi.org/10.1126/science.1161405

Kosfeld, M., Okada, A., & Riedl, A. (2009). Institution formation in public goods games. *The American Economic Review*, *99*(4), 1335–1355. https://doi.org/10.1257/aer.99.4.1335

Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking under the hood of third-party punishment reveals design for personal benefit. *Psychological science*, *27*(3), 405–418. https://doi.org/10.1177/0956797615624469

Krueger, F., & Hoffman, M. (2016). The Emerging Neuroscience of third-party punishment. *Trends in Neurosciences*, *39*(8), 499–501. https://doi.org/10.1016/j.tins.2016.06.004

Kube, S., & Traxler, C. (2011). The interaction of legal and social norm enforcement. *Journal of Public Economic Theory*, *13*(5), 639–660. https://doi.org/10.1111/j.1467-9779.2011.01515.x

Kubota, J. T., Li, J., Bar-David, E., Banaji, M. R., & Phelps, E. A. (2013). The price of racial bias: Intergroup negotiations in the ultimatum game. *Psychological science*, *24*(12), 2498–2504. https://doi.org/10.1177/0956797613496435

Lane, T., Nosenzo, D., & Sonderegger, S. (2023). Law and norms: Empirical evidence. *The American Economic Review*, *113*(5), 1255–1293. https://doi.org/10.1257/aer.20210970

Legros, S., & Cislaghi, B. (2020). Mapping the social-norms literature: An overview of reviews. *Perspectives on Psychological Science*, *15*(1), 62–80. https://doi.org/10.1177/1745691619866455

Lewis, D. (1986). *Convention: A philosophical study*. Wiley.

Li, J., & Zhou, X. (2014). Sex, attractiveness, and third-party punishment in fairness consideration. *PLoS One*, *9*(4), e94004. https://doi.org/10.1371/journal.pone.0094004

Lie-Panis, J., Fitouchi, L., Baumard, N., & André, J. B. (2023). A model of endogenous institution formation through limited reputational incentives.

Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Ensminger, J., Gurven, M., Gwako, E., Henrich, J., Henrich, N., Lesorogol, C., McElreath, R., & Tracer, D. (2008). More 'altruistic'punishment in larger societies. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1634), 587–592. https://doi.org/10.1098/rspb.2007.1517

Marshall, J., & McAuliffe, K. (2022). Children as assessors and agents of third-party punishment. *Nature Reviews Psychology*, *1*(6), 334–344. https://doi.org/10.1038/s44159-022-00046-y

Mathew, S., & Boyd, R. (2011). Punishment sustains large-scale cooperation in Prestate Warfare. *Proceedings of the National Academy of Sciences*, *108*(28), 11375–11380. https://doi.org/10.1073/pnas.1105604108

McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition*, *134*, 1–10. https://doi.org/10.1016/j.cognition.2014.08.013

Miller, D. T., & Prentice, D. A. (1996). The construction of social norms and standards.

Mills, C. W. (2005). "Ideal theory" as ideology. *Hypatia*, *20*(3), 165–183. https://doi.org/10.1111/j.1527-2001.2005.tb00493.x

Molenmaker, W. E., Gross, J., de Kwaadsteniet, E. W., van Dijk, E., & De Dreu, C. K. (2023). Discriminatory punishment undermines the enforcement of group cooperation. *Scientific Reports*, *13*(1), 6061. https://doi.org/10.1038/s41598-023-33167-2

Molho, C., De Petrillo, F., Garfield, Z. H., & Slewe, S. (2024). Cross-societal variation in norm enforcement systems. *Philosophical Transactions of the Royal Society B*, *379*(1897), 20230034. https://doi.org/10.1098/rstb.2023.0034

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. https://doi.org/10.1038/s41467-020-17286-2

Molho, C., & Wu, J. (2021). Direct punishment and indirect reputation-based tactics to intervene against offences. *Philosophical Transactions of the Royal Society B*, *376*(1838), 20200289. https://doi.org/10.1098/rstb.2020.0289

Molleman, L., Kölle, F., Starmer, C., & Gächter, S. (2019). People prefer coordinated punishment in cooperative interactions. *Nature Human Behaviour*, *3*(11), 1145–1153. https://doi.org/10.1038/s41562-019-0707-2

Nobes, G. (1999). Children's understanding of rules they invent themselves. *Journal of Moral Education*, *28*(2), 215–232. https://doi.org/10.1080/030572499103232

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *American Political Science Review*, *86*(2), 404–417. https://doi.org/10.2307/1964229

Pfattheicher, S., Landhäußer, A., & Keller, J. (2014). Individual differences in antisocial punishment in public goods situations: The interplay of cortisol with testosterone and dominance. *Journal of Behavioral Decision Making*, *27*(4), 340–348. https://doi.org/10.1002/bdm.1811

Pfattheicher, S., & Schindler, S. (2015). Understanding the dark side of costly punishment: The impact of individual differences in everyday sadism and existential threat. *European Journal of Personality*, *29*(4), 498–505. https://doi.org/10.1002/per.2003

Piaget, J. (1932/2013). The moral judgment of the child. https://doi.org/10.4324/9781315009681

Pinsof, D. (2023). The evolution of social paradoxes. https://doi.org/10.31234/osf.io/avh9t

Posner, R. A. (1997). Social norms and the law: An economic approach. *The American Economic Review*, *87*(2), 365–369.
Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology*, *67*(4), 741–763. https://doi.org/10.1037//0022-3514.67.4.741
Raihani, N. J., & Bshary, R. (2019). Punishment: One tool, many uses. *Evolutionary Human Sciences*, *1*, e12. https://doi.org/10.1017/ehs.2019.12
Raihani, N. J., Grutter, A. S., & Bshary, R. (2010). Punishers benefit from third-party punishment in fish. *Science*, *327*(5962), 171. https://doi.org/10.1126/science.1183068
Raihani, N. J., Thornton, A., & Bshary, R. (2012). Punishment and cooperation in nature. *Trends in Ecology & Evolution*, *27*(5), 288–295. https://doi.org/10.1016/j.tree.2011.12.004
Rand, D. G., Arbesman, S., & Christakis, N. A. (2011). Dynamic Social Networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences*, *108*(48), 19193–19198. https://doi.org/10.1073/pnas.1108243108
Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M. A. (2009). Positive interactions promote public cooperation. *Science*, *325*(5945), 1272–1275. https://doi.org/10.1126/science.1177418
Rand, D. G., & Nowak, M. A. (2011). The evolution of antisocial punishment in optional public goods games. *Nature Communications*, *2*(1), 434. https://doi.org/10.1038/ncomms1442
Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2012). No third-party punishment in chimpanzees. *Proceedings of the National Academy of Sciences*, *109*(37), 14824–14829. https://doi.org/10.1073/pnas.1203179109
Rockenbach, B., & Milinski, M. (2006). The efficient interaction of indirect reciprocity and costly punishment. *Nature*, *444*(7120), 718–723. https://doi.org/10.1038/nature05229
Ruder, A. I., & Woods, N. D. (2020). Procedural fairness and the legitimacy of agency rulemaking. *Journal of Public Administration Research and Theory*, *30*(3), 400–414. https://doi.org/10.1093/jopart/muz017
Sääksvuori, L., Mappes, T., & Puurtinen, M. (2011). Costly punishment prevails in intergroup conflict. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1723), 3428–3436. https://doi.org/10.1098/rspb.2011.0252
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
Schiller, B., Baumgartner, T., & Knoch, D. (2014). Intergroup bias in third-party punishment stems from both ingroup favoritism and outgroup discrimination. *Evolution and Human Behavior*, *35*(3), 169–175. https://doi.org/10.1016/j.evolhumbehav.2013.12.006
Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological science*, *18*(5), 429–434. https://doi.org/10.1111/j.1467-9280.2007.01917.x
Searle, J. R. (1995). *The construction of social reality*. Simon and Schuster.
Singh, M., Wrangham, R., & Glowacki, L. (2017). Self-interest and the design of rules. *Human Nature*, *28*(4), 457–480. https://doi.org/10.1007/s12110-017-9298-7
Spadaro, G., Liu, J. H., Zhang, R. J., Gil de Zúñiga, H., & Balliet, D. (2023). Identity and institutions as foundations of ingroup favoritism: An investigation across 17 countries. *Social Psychological and Personality Science*, 19485506231172330.
Tan, F., & Xiao, E. (2018). Third-party punishment: Retribution or deterrence? *Journal of Economic Psychology*, *67*, 34–46. https://doi.org/10.1016/j.joep.2018.03.003
Tankard, M. E., & Paluck, E. L. (2017). The effect of a Supreme Court decision regarding gay marriage on social norms and personal attitudes. *Psychological science*, *28*(9), 1334–1344. https://doi.org/10.1177/0956797617709594
Tomasello, M. (2009). *Why we cooperate*. MIT press.
Tyler, T. R. (1994). *The psychology of legitimacy (No. 9425)*. American Bar Foundation.
Tyler, T. R. (1997). Procedural fairness and compliance with the law. *Revue Suisse d'Economie Politique et de Statistique*, *133*, 219–240.
Tyler, T. R. (2003). Procedural justice, legitimacy, and the effective rule of law. *Crime and justice*, *30*, 283–357. https://doi.org/10.1086/652233
Tyler, T. R., & Boeckmann, R. J. (1997). Three strikes and you are out, but why? The psychology of public support for punishing rule breakers. *Law & Society Review*, *31*(2), 237–265. https://doi.org/10.2307/3053926
Tyler, T. R., & Lind, E. A. (1992). A relational model of authority in groups. *Advances in Experimental Social Psychology*, *25*, 115–191. Academic Press. https://doi.org/10.1016/s0065-2601(08)60283-x
Van Bavel, J. J., Pärnamets, P., Reinero, D. A., & Packer, D. (2022). How neurons, norms, and institutions shape group cooperation. *Advances in Experimental Social Psychology*, *66*, 59–105. Academic Press. https://doi.org/10.1016/bs.aesp.2022.04.004

Weber, M. (1921). *Economy and society*. Scribner and Sons.
Weidman, A. C., Sowden, W. J., Berg, M. K., & Kross, E. (2020). Punish or protect? How close relationships shape responses to moral violations. *Personality and Social Psychology Bulletin*, *46*(5), 693–708. https://doi.org/10.1177/0146167219873485
Whitley, B. E. (1999). Right-wing authoritarianism, social dominance orientation, and prejudice. *Journal of Personality and Social Psychology*, *77*(1), 126–134. https://doi.org/10.1037/0022-3514.77.1.126
Wu, J., Balliet, D., & Van Lange, P. A. (2016). Gossip versus punishment: The efficiency of reputation to promote and maintain cooperation. *Scientific Reports*, *6*(1), 23919. https://doi.org/10.1038/srep23919
Wu, J.-J., Zhang, B.-Y., Zhou, Z.-X., He, Q.-Q., Zheng, X.-D., Cressman, R., & Tao, Y. (2009). Costly punishment does not always increase cooperation. *Proceedings of the National Academy of Sciences*, *106*(41), 17448–17451. https://doi.org/10.1073/pnas.0905918106
Wylie, J., & Gantman, A. (2023). Doesn't everybody jaywalk? On codified rules that are seldom followed and selectively punished. *Cognition*, *231*, 105323. https://doi.org/10.1016/j.cognition.2022.105323
Wylie, J., Milless, K. L., Sciarappo, J., & Gantman, A. (2024). The biased enforcement of rarely followed rules. *Personality and Social Psychology Bulletin*, 01461672241252853. https://doi.org/10.1177/01461672241252853
Yang, H., Zhang, Y., Lyu, Y., & Tang, C. (2023). Group bias under uncertain environment: A perspective of third-party punishment. *Acta Psychologica*, *237*, 103957. https://doi.org/10.1016/j.actpsy.2023.103957
Yankah, E. N. (2024forthcoming). Deputization and privileged white violence. *Stanford Law Review*.
Young, H. P. (2015). The evolution of social norms. *Economics*, *7*(1), 359–387. https://doi.org/10.1146/annurev-economics-080614-115322
Yudkin, D. A., Rothmund, T., Twardawski, M., Thalla, N., & Van Bavel, J. J. (2016). Reflexive intergroup bias in third-party punishment. *Journal of Experimental Psychology: General*, *145*(11), 1448–1459. https://doi.org/10.1037/xge0000190

AUTHOR BIOGRAPHIES

Jordan Wylie is an NSF SBE Postdoctoral Research at Boston College. She received her PhD in 2022 from CUNY. Her research examines the influences of morality, norms, and rules on judgments and decision-making, including how moral information shapes how we reason about rules and rule breakers, and when we are curious to learn more about them. She has published 10+ peer-reviewed papers and chapters. Dr. Wylie has received an Outstanding Research award and a Jenessa Shapiro Graduate Research Award from the Society for Personality and Social Psychology.

Ana Gantman is an Assistant Professor of Psychology at Brooklyn College and the CUNY Graduate Center. She has published 20+ peer-reviewed papers and chapters which have been cited ~600 times ($h = 12$, i10 = 11). Dr. Gantman's lab investigates how our moral values tune our understanding of the law, public policy, and our curiosity and attention. Dr. Gantman received the SAGE Young Scholars Award from the Society for Personality and Social Psychology and was named an APS Rising Star.

How to cite this article: Wylie, J., & Gantman, A. (2024). Cooperation, domination: Twin functions of third-party punishment. *Social and Personality Psychology Compass*, e12992. https://doi.org/10.1111/spc3.12992