Resentment and Punishment[*]

Erik O. Kimbrough[†] **Alexander Vostroknutov**[‡ §]

April 26, 2023

Abstract

Punishment is understood to be necessary to the creation and maintenance of social norms. We argue that it can be driven by *resentment* that arises when norms are violated. Using a model of endogenous norms, we show how to define a norm-violation on any game or choice set and how to predict which actions will generate resentment and thus be the target of punishment. We analyze evidence from previous experiments and show that our model of resentment-driven punishment can explain observed second- and third-party punishment decisions across diverse settings.

JEL classifications: C91, C92
Keywords: norms, social preferences, experiments, punishment

[*] This paper contains material drawn from an earlier working paper, titled "A Theory of Injunctive Norms". We would like to thank Valerio Capraro, Hartmut Kliemt, Sébastien Pouget, David Rojo-Arjona, Nicole Saito, Matthias Stefan, Elias Tsakas, and participants in the Works-in-Progress seminar at Chapman University's Smith Institute for Political Economy and Philosophy, the ESA-World meetings in Vancouver, the PPE Summer School at Chapman University, and seminars at GSBE-ETBC Maastricht, Ohio State, George Mason, Konstanz, Max Plank Institute-Bonn, UC-Davis, Virginia Tech, and San Jose State for insightful comments and suggestions. All mistakes are our own.

[†] Smith Institute for Political Economy and Philosophy, Chapman University, One University Drive, Orange, CA 92866, USA. email: ekimbrou@chapman.edu.

[‡] Department of Economics (MPE), Maastricht University, Tongersestraat 53, 6211LM Maastricht, The Netherlands. e-mail: a.vostroknutov@maastrichtuniversity.nl.

[§] Corresponding author.

1 Introduction

Recent models propose that other-regarding and context-dependent behavior can be parsimoniously explained if people care not only about consumption utility but also about the extent to which their actions conform to injunctive norms that define what one *ought* to do in a particular setting (Bicchieri, 2006; López-Pérez, 2008; Cappelen et al., 2007; Kessler and Leider, 2012; Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2016). Injunctive norms are often prosocial and aimed at cooperation, such that following them is welfare enhancing. However, since motives mix both norm-following and self-interest, these models imply that individual decisions will rarely result in full norm compliance. In practice, this means that the descriptive norm (what people actually do) need not coincide with the injunctive norm. To ensure that injunctive and descriptive norms coincide, another ingredient is needed: punishment.

It is well-established that creating and sustaining norms often requires punishment of violations and that without punishment mechanisms, the evolution of social norms and of preferences for following them would be impossible (Kandori, 1992; Chudek and Henrich, 2011; Henrich, 2015). In the realm of incentivized economic behavior, the concept of "altruistic punishment" proposed by Fehr and Gächter (2002) reflects this idea: people punish norm violators for the sake of encouraging norm adherence and will even incur punishment costs without regard for personal benefit in the form of future payoffs, reputation, etc. We argue that such punishment can be driven by *resentment* of actions that violate injunctive norms.

Kimbrough and Vostroknutov (2023) provide an account of endogenous norms, in which the normatively best outcome in a given choice set is the one that minimizes the aggregated "dissatisfaction" of all interested parties, and outcomes are ranked according to how much dissatisfaction they generate.

In the model, individual dissatisfaction is derived from self-interest, as individuals compare the utility they actually receive at any outcome to the counterfactual possibilities that would yield them a higher utility. Aggregate dissatisfaction is then derived from empathy, with individuals recognizing that others feel similar dissatisfaction when outcomes leave them with less than they could have received. The normatively best outcome (aka the injunctive norm), then, is defined as a sort of compromise between the interests of agents that minimizes the (possibly weighted) sum of agents' dissatisfaction.

The model takes individual preferences over consumption and the set of feasible outcomes as given and then defines norms for the given choice set. However, the extent to which behavior conforms to norms in the model depends largely on individuals' propensities to adhere to norms. Individuals with sufficiently high norm-following propensity may follow norms fully, but others' choices will be modulated by self-interest. This leaves a potential gap between the injunctive norm that tells people what they ought to do, and the descriptive norm of actually common behavior. A model of norms is incomplete without a punishment mechanism that maintains norm compliance by deterring selfish urges to gain more payoff at the expense of breaking the norm.

Here we argue that the model introduced by Kimbrough and Vostroknutov (2023) also implicitly defines the set of actions that would generate resentment and thus be the target of punishment. We assume that actions generate resentment to the extent that they cause the outcome to deviate from the norm. Such resentment can be seen as a motive for a meta-norm of punishment and so can help reduce the gap between injunctive and descriptive norms.

We build upon the model of context-dependent norms, to define a model of context-dependent punishment. The model assumes that players have norm-dependent utility functions that trade off consumption utility against adherence to shared injunctive norms. Agents are defined by their consumption utility and their propensity to follow norms, and the same propensity motivates their willingness to punish violations thereof. A violation is detected when an agent takes an action that renders the normatively best outcome infeasible, and the magnitude of the violation is proportional to the gap between the most appropriate remaining feasible outcome and the most appropriate outcome in the original choice set. Agents' *resentment* is a measure of the magnitude of the norm violation.

Since punishment is itself a norm, this requires a notion of appropriate punishment, which is assumed to satisfy two principles: deterrence and proportionality. Deterrence requires that the most appropriate level of punishment is sufficient to deter a norm violation, and proportionality requires that the cost of punishment to the violator is proportional to the resentment of the enforcer.

The model can be applied to any game with observable actions by recursively computing punishment costs for both enforcer and violator and integrating these into the utility function over the terminal nodes of the game. The transformed game can be analyzed with standard game theoretic tools. We consider two ways of implementing punishment in games: 1) punishment enacted via actions that harm the norm violator within the game itself (e.g. A rejects B's offer in an ultimatum game) and 2) punishment enacted by a separate enforcement mechanism (e.g. A can pay to reduce B's payoff after observing B's decision in a dictator game).

To illustrate, we analyze second-party punishment in a set of games due to Charness and Rabin (2002) where agents must take actions within the game to enact punishment, and we analyze both second- and third-party punishment in games introduced by Fehr and Fischbacher (2004) that include a separate punishment mechanism. We highlight how the model makes predictions about which actions constitute violations and hence which actions ought to be the target of punishment (and how severely they ought to be punished). We show that the model can readily account for the observed punishment by reference to resentment of violations of context-dependent norms.

2 Model

2.1 Injunctive Norms

Kimbrough and Vostroknutov (2023) present an axiomatic model that defines injunctive norms in terms of shared judgments about what is socially appropriate and inappropriate, and they argue that such judgments give each action a "normative valence". In their model, the normative valence of any outcome depends only on the set of feasible outcomes and does not depend on a game's strategic structure defined by a sequence of moves, information sets, etc. Take a set C of *outcomes* with $|C| > 1$ and a finite set of players N (Osborne and Rubinstein, 1994). Let $u : C \to \mathbb{R}^N$ be a utility function (synonymous with payoff function) that assigns to each outcome a vector of players' utilities (payoffs) with $u_i(x)$ meaning the payoff of player i for outcome $x \in C$.

The normative valence of an outcome is defined in terms of comparative *dissatisfaction*, with the normatively most appropriate outcome in the feasible set being the dissatisfaction-minimizing outcome. Each individual evaluates the dissatisfaction that they would feel at each feasible outcome by comparing it to each element in the set of other feasible outcomes that are preferable. The total dissatisfaction with a particular outcome x is then the sum of the dissatisfaction derived from all other feasible outcomes that yield a higher utility.

The main ingredient of the model is

$$d_i(x, c) := \max\{u_i(c) - u_i(x), 0\}, \tag{1}$$

the *dissatisfaction* that player i feels about outcome x *because of* the possibility of another outcome c. This notion of dissatisfaction is intended to capture attention to counterfactuals. If the outcome is x, then player i suffers dissatisfaction from it to an extent $d_i(x, c)$ because c could have been the outcome instead. Dissatisfaction is positive when c brings player i more utility than x and zero otherwise.

A player's total dissatisfaction with an outcome then depends on the counterfactual comparison to *all* feasible outcomes is

$$D_i(x \mid C) := \int_{c \in C} d_i(x, c) dc. \tag{2}$$

An agent is more dissatisfied with an outcome when there are more and better alternatives.

Finally, the *aggregate dissatisfaction* of x—that is, dissatisfaction aggregated across all players—is

$$D(x \mid C) := \sum_{i \in N} D_i(x \mid C). \tag{3}$$

This aggregation is intended to capture empathy, with individuals applying their knowledge of how others would feel at any outcome to agree upon a normative ranking.

Finally, Kimbrough and Vostroknutov (2023) assume that the normative valence of x is inversely proportional to its aggregate dissatisfaction, such that the normatively best outcome is the one that minimizes aggregate dissatisfaction. Let $\langle N, C, u, D \rangle$ be an *environment*, and consider the following definition:

Definition 1. *For an environment $\langle N, C, u, D \rangle$, call $\eta : C \to [-1, 1]$, defined as*

$$\eta(x \mid C) := [-D(x \mid C)]$$

*where $[\cdot]$ is the linear normalization of $-D$ to $[-1, 1]$, a **norm function** associated with $\langle N, C, u, D \rangle$. If D is a constant function, set $\eta(x) = 1$ for all $x \in C$.*

The norm function η is the negative of aggregate dissatisfaction, normalized to the interval $[-1, 1]$. Thus, the outcome x with $\eta(x|C) = 1$ is the most socially appropriate (the norm) and the one with $\eta(x|C) = -1$, the least socially appropriate.

2.2 Resentment

Defining a norm in this way also leads to a natural definition of a norm violation. We assume that individuals who observe norm violations by others feel *resentment* and that this resentment undergirds a second norm: one ought to punish norm violators. We use this idea to model normative reaction to any action by a player that makes it impossible to achieve the most socially appropriate feasible outcome. Punishment mechanisms have been modeled before in theories of reciprocity (see e.g., Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). However, our approach is different in that we conjecture that a meta-norm (of punishment for violations) is activated when the primary norm is not followed in a particular setting.[1]

We start with an environment $\langle N, C, u, D \rangle$ that defines the normative valence of all outcomes and assume that player i chooses one action from some set A. Each choice $a \in A$ restricts the set of reachable outcomes to $C_a \subseteq C$, but it does not change the norm function η since player i made her choices taking all possible outcomes into account. Let $M = \text{argmax}_{x \in C}\, \eta(x|C)$ be the set of the most socially appropriate outcomes, and suppose that action a is chosen so that $C_a \cap M = \emptyset$. In other words, player i has chosen an action that makes all most socially appropriate outcomes unreachable. We define such a choice as a norm violation. The *resentment* of this violation can be defined as the difference between the normative valence of the most socially appropriate outcome $\max_{x \in C} \eta(x|C)$, which has been rendered unreachable, and the most socially appropriate

[1] Kimbrough and Vostroknutov (2016) provide evidence in support of this conjecture: they find that subjects with high/low propensity to adhere to norms also have high/low rejection thresholds in the Ultimatum game (UG). This result suggests that punishment in the UG is normative in nature.

outcome that can still be obtained after a was chosen ($\max_{x \in C_a} \eta(x|C)$).[2] Denote this difference by

$$r_a := \max_{x \in C} \eta(x|C) - \max_{x \in C_a} \eta(x|C).$$

Since the model of injunctive norms provided by Kimbrough and Vostroknutov (2023) defines the appropriateness of all outcomes in any game or choice set endogenously, our definition of resentment also delineates the set of actions that merit punishment for any game or choice set.

The next step is to construct a model that details how player i, who violated the norm, should be punished. In what follows we propose a specific model of punishment that can be applied to make more specific testable predictions not only about who will be the target of punishment and when but also how much punishment they might expect to receive.

2.3 Punishment

Punishment is a complex phenomenon, and there might be many reasons for it: the desire to achieve the most appropriate outcome, revenge, reputation concerns, etc. It is not our goal here to capture all these motives, as the empirical evidence on their relative prominence is, at best, scarce. Here we concentrate on punishment that is norm-driven, and we model it by reference to two core principles on which punitive systems seem to be based. One is the *deterrence* principle, which states that the amount of punishment should be large enough that a player does not have an incentive to violate the norm. Another is *proportionality* principle, which states that the amount of punishment should be proportional to r_a, the degree of norm violation (as in "an eye for an eye").

Since punishment is, itself, a norm in this framework, making predictions about the magnitude of punishment requires a model that determines the normative valence of each possible punishment decision, in a given setting. To construct the normative valences pertaining to punishment we determine, for each possible payoff of the norm violator i, how "punishment-appropriate" it would be, given that she chose action a (with $r_a > 0$). We highlight three important elements: 1) the payoff that i would have gotten in the most socially appropriate outcome, $u_{im} = \max_{x \in M} u_i(x)$, or the payoff that she chose to forgo when choosing a; 2) the minimal payoff that she can obtain in the game, $\underline{u}_i = \min_{x \in C} u_i(x)$, which serves as a reference point for the harshest punishment possible; and 3) the payoff that i seemingly "aimed at" receiving after

[2] In defining the degree of norm violation in this way we make an implicit assumption that after a was chosen there remains a consensus among players that the outcome with normative valence $\max_{x \in C_a} \eta(x|C)$ remains reachable. This "optimistic" scenario is by no means the only way the degree of violation could be perceived. However, whether this is so, or whether the degree of violation is calculated differently, is an empirical question that we do not try to answer in this paper and instead leave for future experimental investigations (one study that tests our model of punishment is Merguei et al. (2022)). What is important is that the degree of violation is weakly monotonic in r_a (defined below).

choosing a, $\bar{u}_{ia} = \max_{x \in C_a} u_i(x)$.[3] The deterrence principle says that it is very inappropriate for i to receive a payoff that exceeds $m = \min\{u_{im}, \bar{u}_{ia}\}$. In most cases $\bar{u}_{ia} > u_{im}$, thus, after i chose action a, we assume that the punishment norm function dictates that she not enjoy a payoff higher than the one she would have received if she followed the primary norm (i.e., her payoff at the most socially appropriate outcome).[4] The proportionality principle requires that punishment should be proportional to r_a, with the harshest punishment—reducing i's payoff to its minimum $\underline{u}_i$—being applied when the norm is violated to the fullest extent ($r_a = 2$). We propose a punishment norm function $\mu_i : \mathbb{R} \to [-1, 1]$, which is a mapping from violator i's payoffs to the normative valence space shown on Figure 1.

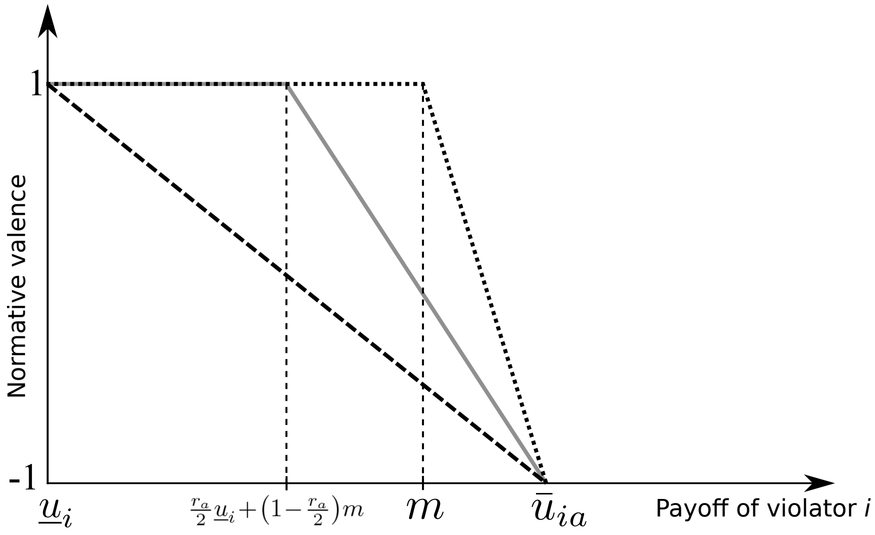

Figure 1: Punishment norms for $r_a < 2$ (solid gray line), $r_a = 2$ (dashed line), and $r_a \to 0$ (dotted line).

Notice first that μ_i is defined for all payoffs on the interval $[\underline{u}_i, \bar{u}_{ia}]$ from the lowest possible payoff in the whole game to the maximum payoff that remains achievable after a. The properties of μ_i are as follows. The payoff $\bar{u}_{ia}$, which constitutes i's "likely intent" (that is, the payoff we assume was the aim of the norm violation) has the lowest possible normative valence of -1, and all payoffs less than that have higher normative valence. Next, note that social appropriateness reaches its maximum when the payoff drops to $\frac{r_a}{2}\underline{u}_i + \left(1 - \frac{r_a}{2}\right)m$. This value is linearly proportional to r_a and is equal to m when $r_a \to 0$ and to $\underline{u}_i$ when $r_a = 2$. This point is calculated by applying the proportionality principle and represents the maximum appropriate punishment proportional to r_a, taking into account the deterrence principle (which imposes the constraint that the punishment should not be less than m). All payoffs less than $\frac{r_a}{2}\underline{u}_i + \left(1 - \frac{r_a}{2}\right)m$ have the

[3] Of course, in a game it is not obvious that player i can guarantee herself payoff $\bar{u}_{ia}$, since other players might move in the subgame. However, we follow a long tradition, going at least back to Elster (1989), in assuming that normative thinking, of which punishment is an example, is not strategic. In law practice, criminal intent is reason enough for punishment regardless of the plausibility of achieving the intended outcome.

[4] In rare cases in which $\bar{u}_{ia} \leq u_{im}$, we assume the punishment norm implies that player i should still be punished for norm violation by having even less than $\bar{u}_{ia}$.

highest normative valence of 1. This implies that for the punishers it is normatively irrelevant whether i gets punished by having payoff $\frac{r_a}{2}\underline{u}_i + (1 - \frac{r_a}{2})m$ or lower.

2.4 Punishment Mechanisms

Punishment can be implemented in two different ways. The first, which is perhaps the most natural way, is to punish "outside the game." This requires the existence of a separate punishment mechanism that allows players to decrease each others' payoffs *without deviating from the normatively appropriate actions defined by the game itself*. This is exactly the idea that is widely used today in experimental economics since Fehr and Gächter (2000), who introduced a punishment technology to the repeated Public Goods game. Indeed, such a mechanism makes it possible to achieve two normative goals that we assume agents have: they can reach the most socially appropriate outcome remaining in the subgame after a was chosen, and they can *separately* punish player i for the norm violation. If such a punishment mechanism exists, then the punishment function is defined by $\sigma + (1-\sigma)\mu_i(p)$ for payoff p of player i. The parameter $\sigma \in [0,1]$ represents the relative importance of punishment in a given situation. When $\sigma = 1$ all punishment options are equally (and maximally) socially appropriate; thus, the least costly punishment will be chosen. When $\sigma = 0$, the players feel that punishment is most important. We discuss punishment mechanisms in much more detail in Section 3.

The second way to implement punishment is by taking action within the game itself (e.g., when an outside-the-game punishment mechanism is not available). This leads to an additional complication in standard games, which do not assume punishment mechanisms: players are forced to combine the main normative goal of the game and punishment in one normative space. We assume that they combine these normative motivations by taking a convex combination of the norms η and μ_i, thereby, increasing the normative valence of the outcomes that decrease i's payoff. Abusing notation, let us think of the function μ_i, originally defined on the space of payoffs, as a function defined on outcomes with $\mu_i(x)$ meaning $\mu_i(u_i(x))$ and assume that for each $x \in C_a$ the combined norm function is

$$\eta'(x|C) = \sigma\eta(x|C) + (1-\sigma)\mu_i(x).$$

Here, again, the parameter σ defines the relative importance of punishment. To illustrate how this amalgamation of norms works we analyze the Ultimatum game in Example 1 (Güth et al., 1982). The intuition is that normatively inappropriate offers by the proposer can "justify" (in the sense of generating norm-driven resentment) retaliatory rejection by responders.

Example 1. Ultimatum Game (UG) with punishment norm. From the perspective of our framework, the UG can be viewed as a dictator game with a rather extreme punishment mechanism, which only allows maximal punishment. The set of outcomes is $C = [0,1] \cup \{p_c \mid c \in [0,1]\}$ with utilities $u(c) = (1-c, c)$ if the offer is accepted and $u(p_c) = (0,0)$ for all $c \in [0,1]$ if the offer is rejected. Here outcome p_c represents rejection in the subgame that follows the choice of c.

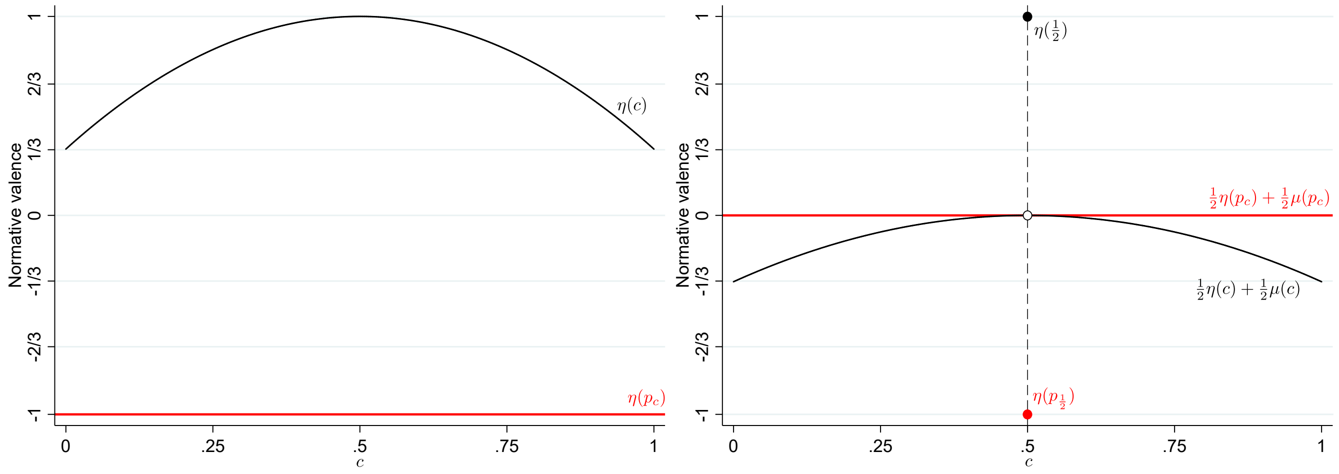

Figure 2: Norm functions in UG. **Left:** the norm function. **Right:** for each $c \in C$, except $c = \frac{1}{2}$ which is the norm, convex combination of punishment function and norm function.

The left graph on Figure 2 shows the norm function in the UG. The black line corresponds to the accepted divisions $(1-c, c)$ and the red line to the rejection outcomes p_c. The right graph shows, for each proposal $c \in [0,1]$, the norm function from the left graph combined with the punishment function μ. This function is defined, as described in Section 2.3, to be $\mu(c) = -1$ (since the payoff $1-c$ is the highest attainable payoff for the proposer in the subgame following offer c) and $\mu(p_c) = 1$ for all $c \neq \frac{1}{2}$ (since p_c is the outcome with the lowest possible payoff for the proposer). Notice that for any deviation from the equal split, the normative valence of rejecting p_c is higher than the normative valence of accepting the offer $(1-c, c)$. The difference increases as the offer yields increasingly unequal divisions. Intuitively, this helps explain why more unequal offers are more likely to be rejected (see e.g. Güth and Kocher, 2014, for a survey of evidence). □

3 Norms and Punishment in Games with Observable Actions

In this section we put all the elements of our model together and analyze how extensive and normal form games with norm-dependent utility and punishment are played.

We start by defining a utility function that takes injunctive norms as an input. Up to this point our model was purely normative, in the sense that it only described how appropriate or inappropriate the outcomes of actions can be. However, we never talked about the actual goals of the players. The last, very important, ingredient that is still missing is the consumption utility that players enjoy from receiving their payoffs. We follow previous studies (Kessler and Leider, 2012; Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2016) and define player i's *norm-*

dependent utility of outcome x in a game as:

$$w_i(x) := u_i(x) + \phi_i \eta(x),$$

where $u_i(x)$ is the consumption utility of outcome x, with the set of outcomes C corresponding to the set of terminal nodes in the game. For ease of exposition, we replace the notation $\eta(x|C)$ with $\eta(x)$ since C is always the same when analyzing a particular game. $\phi_i \geq 0$ is a constant that defines player i's norm-following propensity (Kimbrough and Vostroknutov, 2016, 2018; Tate et al., 2022). This last parameter defines how important following norms is for player i: if $\phi_i = 0$ we have a standard utility maximizer, as $\phi \to \infty$ we have player i who only cares about following norms.

Let a tuple $\Gamma = \langle N, C, u, D, H \rangle$ be an *extensive form game with observable actions* with the set of outcomes C corresponding to the set of terminal nodes and H being the finite set of histories. Notice that Γ is a standard game with utilities being the material payoffs or consumption utilities.

Let us define some notation. $h = (a^1, a^2, ..., a^\ell)$ represents a history of length ℓ where $a^t = (a_1^t, ..., a_N^t)$ is a profile of actions chosen at stage t, $1 \leq t \leq \ell$. Each history h becomes commonly known to all players once it occurs. Empty history $\{\varnothing\} \in H$ represents the beginning of the game. After history h player i has the set of actions $A_i(h)$, which is empty if and only if $h \in C \subsetneq H$, where C is thought of as the set of all terminal histories. Let $p(h)$ denote the history immediately preceding h and C_h the set of terminal nodes that can occur after h.

3.1 Games without a Separate Punishment Mechanism

We first consider how punishment works in the absence of a specially defined punishment mechanism. Our goal is to define the norm function at each history and to determine the norm-dependent utilities in the terminal nodes. We proceed recursively and define the norm function at history h, which is a function $\eta^h : C_h \to [-1, 1]$ that attaches normative valences to all outcomes following h, through the norm function in the immediately preceding history $p(h)$ and a punishment function. Notice that the norm function at the beginning of the game is defined as in Section 2.1. Namely, $\eta^{\{\varnothing\}} = \eta$.

We assume that at any history h the players reason *locally* about the changes in the norm function that need to be made. Specifically, they take $\eta^{p(h)}$ and reason about who should be punished for the actions taken in $p(h)$ that led to h. This means that they combine $\eta^{p(h)}$ with the punishment functions $\mu_i^{a_i}$ where $a_i \in A_i(p(h))$ is the action of player i that led to h. Thus, to determine η^h we need to specify punishment functions $\mu_i^{a_i}$ and the way they are combined with $\eta^{p(h)}$.

To define $\mu_i^{a_i}$ we use the same logic as in Section 2.3. We determine the degree of norm violation of each player i and construct the punishment in the i's payoff space. Let $C_{p(h)}^{a_i} \subseteq C_{p(h)}$ be the set of outcomes reachable given the choice a_i of player i. Notice that $C_{p(h)}^{a_i}$ is weakly larger than C_h, the set of outcomes reachable in h, since players choose in a normal-form stage game and the actions of other players are not restricted. It makes sense to consider $C^{a_i}_{p(h)}$ as a set of outcomes that should be used for the determination of punishment since player i cannot be held responsible for what other players choose. Let $M_{p(h)} = \operatorname{argmax}_{x \in C_{p(h)}} \eta^{p(h)}(x)$ be the set of the most appropriate outcomes according to $\eta^{p(h)}$. In the simplest case, all players choose actions a_i that leave some outcomes in $M_{p(h)}$ reachable. If this happens, then no one should be punished and the norm function in h is the same as the norm function in $p(h)$. In other words, set

$$\eta^h = \eta^{p(h)} \quad \text{if } \forall_{i \in N} \;\; C^{a_i}_{p(h)} \cap M_{p(h)} \neq \emptyset.$$[5]

Here the understanding is that η^h is equal to $\eta^{p(h)}$ on its domain, which is the subset of the domain of $\eta^{p(h)}$. If at least one player has chosen an action which makes all outcomes in $M_{p(h)}$ unreachable then players go into the "punishment mode" in which the punishment functions are combined with the original $\eta^{p(h)}$. To determine $\mu^{a_i}_i$ we first calculate the degree of norm violation for player i as

$$r^{a_i}_i = \max_{x \in C_{p(h)}} \eta^{p(h)}(x) - \max_{x \in C^{a_i}_{p(h)}} \eta^{p(h)}(x).$$

$r^{a_i}_i$ is positive only for players who chose the actions inconsistent with all outcomes in $M_{p(h)}$. Let $V = \{i \mid r^{a_i}_i > 0\} \subseteq N$ denote the set of such players. For each $i \in V$ we define three payoffs: 1) the payoff that i would have gotten in the most socially appropriate outcome, $u_{im} = \max_{x \in M_{p(h)}} u_i(x)$, or the payoff that she chose to forgo when choosing a_i; 2) the minimal payoff that she can obtain in the whole game, $\underline{u}_i = \min_{x \in C} u_i(x)$, which serves as a reference point for the harshest punishment;[6] and 3) the payoff that i "aims at" by choosing a_i, $\bar{u}_i = \max_{x \in C^{a_i}_{p(h)}} u_i(x)$. Let $m_i = \min\{u_{im}, \bar{u}_i\}$ and define the punishment norm function $\mu^{a_i}_i$ as shown in Figure 1 in Section 2.3. Finally we calculate the norm function η^h by combining $\eta^{p(h)}$ and the punishment functions $(\mu^{a_i}_i)_{i \in V}$:

$$\eta^h(x) = \sigma \eta^{p(h)}(x) + (1-\sigma) \frac{\sum_{i \in V} \mu^{a_i}_i(x)}{|V|} \quad \forall_{x \in C_h}$$

where $\mu^{a_i}_i(x)$ is short for $\mu^{a_i}_i(u(x))$. Essentially, η^h is a convex combination of $\eta^{p(h)}$ and the average punishment function that gives equal weights to all players.[7]

The construction above shows how to calculate the norm function for each node in game Γ. Since the norm function at the beginning of the game is known to be $\eta^{\{\emptyset\}} = \eta$, we can

[5] Notice that this definition allows for the possibility that each player chooses the action consistent with some outcome in $M_{p(h)}$, but the resulting action profile $a = (a_i)_{i \in N}$ makes all outcomes in $M_{p(h)}$ unreachable.

[6] An alternative possibility is to consider history dependent punishment reference points $\underline{u}_i(h) = \min_{x \in C_h} u_i(x)$. We leave it to the future research to determine whether the harshest punishment options are perceived as history dependent or constant.

[7] Alternative definitions are possible. For example, instead of the average punishment function, a more punishment oriented approach would be to take the envelope of the punishment functions $\max\{\mu^{a_1}_1(x), ..., \mu^{a_N}_N(x)\}$.

recursively compute the norm functions for all histories $h \in H \backslash C$. The last step is to redefine the payoffs in Γ with the norm-dependent utility. Let $\Gamma' = \langle N, C, w, D, H \rangle$ be the same game only with utilities defined by

$$w_i(x) := u_i(x) + \phi_i \eta^{p(x)}(x) \quad \forall_{i \in N} \forall_{x \in C}$$

where $\eta^{p(x)}$ is the norm function in the node that immediately precedes terminal node x. Γ' is a standard extensive form game that can be analyzed using any equilibrium concept.

3.2 Games with a Separate Punishment Mechanism

In the previous section we showed how to introduce norms into any game with observable actions without separate punishment mechanisms. Most games analyzed in the literature fall under this category. However, this construction also carries certain implicit assumptions. For example, the fact that punishment functions are amalgamated into the norm function of the game as it unfolds implies that punishment for a single act of "wrong-doing" at history h has influence on all subsequent histories and eventually final payoffs. In other words, the model above has no absolution, which entails that violators are punished for each norm violation until the end of the game. This might not be the most realistic way in which punishment is actually carried out. If an external punishment mechanism exists "outside" of the game, it is reasonable to think that each norm violation is punished with this mechanism right after it occurs, and that this punishment absolves the violation. This latter point implies that there is no need to update the norm function in the game itself and it proceeds in accordance with the original norm function defined before the game started.[8]

In this section we show how to incorporate norms assuming that punishment can be exercised outside the game. We start with the same game Γ as before and the norm function η defined for it. We assume that as the game is played there is a possibility for each player to punish any other player at each history $h \in H$. Notice that this includes the terminal nodes C, which means that punishment can be carried out after the last move in the game as well. The norm function η in the game stays unchanged, so players receive norm-dependent utility in accordance with it. In addition, the final payoffs are adjusted with the costs of punishment that players incur and the punishment that they receive from other players.

We set up the punishment mechanism as follows. As before suppose we are at history h and the actions $a_i \in A_i(p(h))$ for all $i \in N$ are those that lead to h. We determine the punishment functions $\mu_i^{a_i}$ in the same way as in the previous section only with $\eta^{p(h)} = \eta$ on its domain. $\mu_i^{a_i}$

[8] Though, it should be noted that the model without a separate punishment mechanism does have one desirable property: players who do not want to punish others get punished themselves, since the updated norm function in each history incorporates the punishment. The model with a separate punishment mechanism, at least the way we put it, does not have this property.

is a function from the payoff interval $[\underline{u}_i, \bar{u}_i]$ to normative valences $[-1, 1]$. Assume that each player $j \neq i$ has access to a punishment mechanism that allows j to decrease i's payoff with a cost. Suppose that j believes that without punishment i will get her desired payoff $\bar{u}_i$, so j solves the following maximization problem to decide how much payoff to subtract from i:

$$s_{ji}^{a_i} = \arg \max_{s \in [0, \bar{u}_i - \underline{u}_i]} \phi_j(\sigma + (1-\sigma)\mu_i^{a_i}(\bar{u}_i - s)) - \zeta(s).$$

Here $\phi_j \geq 0$ is j's norm-following propensity; $\sigma + (1-\sigma)\mu_i^{a_i}(\bar{u}_i - s)$ is the punishment norm function adjusted with the weight σ as in Section 2.3; and $\zeta(x)$ is an increasing cost function with $\zeta(0) = 0$.[9] $s_{ji}^{a_i}$ is the amount of payoff that j has decided to subtract from i. Let $q_{ji}^{a_i} = \zeta(s_{ji}^{a_i})$ denote the cost that j incurs for the punishment of i. This essentially defines the costs that j and i have from punishment.[10]

Notice that the punishment decisions are not strategic and happen separately from the game. The way that the players take the punishment into account is through the losses they suffer at the end of the game. We redefine the payoffs in Γ by considering a modified game $\Gamma'' = \langle N, C, v, D, H \rangle$ with utility for player i calculated as follows. For any outcome $x \in C$, which is also a terminal history, let us write $x = (a^1, a^2, ..., a^\ell)$, where $a^t = (a_1^t, ..., a_N^t)$ is the action profile chosen in stage t that leads to x. Let

$$v_j(x) := u_j(x) + \phi_j \eta(x) - \sum_{t=1}^{\ell} \sum_{i \neq j} q_{ji}^{a_i^t} + s_{ij}^{a_j^t}.$$

The utility v_j is simply the norm-dependent utility with the norm function η minus the punishment that player j incurs on the way to x and the cost of punishment that j metes out to others. Γ'' is a standard extensive form game that can be solved by any equilibrium concept.

It should be noted that the way we construct Γ'' has many ad hoc assumptions about how exactly punishment is done. There are a plethora of variants that can be considered, but we leave alternative ways of constructing the punishment norm for future work.

4 Evidence

In this section we test our theory of punishment for norm violations. We show that the model can account for behavior in games where the first mover does not choose the most appropriate outcome, and the model predicts that this behavior should be punished. We compare the model's predictions to the experiments of Charness and Rabin (2002) who study several games

[9] In the experiments subjects usually pay one experimental unit to subtract three from the punished player. In this case $\zeta(x) = \frac{x}{3}$.

[10] By the definition of μ_i in Section 2.3, if no violation of the norm happened then $\mu_i = 1$. In this case j optimally chooses to not punish, or pay 0 for it.

that allow for the possibility of punishment by the second-mover. Then we consider third-party punishment games due to Fehr and Fischbacher (2004) who employ a formal outside-the-game punishment mechanism and show that punishment decisions are consistent with those implied by the model.

Case 1. Charness and Rabin (2002). The authors (CR) study the games shown in Figure 3 that neatly illustrate how punishment works in our model when there is no external punishment mechanism. The games A1 and A2 are identical except for the payoffs that the players get if P1 ends the game with the first move ((750, 0) vs. (550, 550)). The same is true for the games B1 and B2. The A and the B games also differ in that, in the A games, P2 has a material incentive to choose (400, 400); whereas, in the B games, P2 is materially indifferent between the two options.

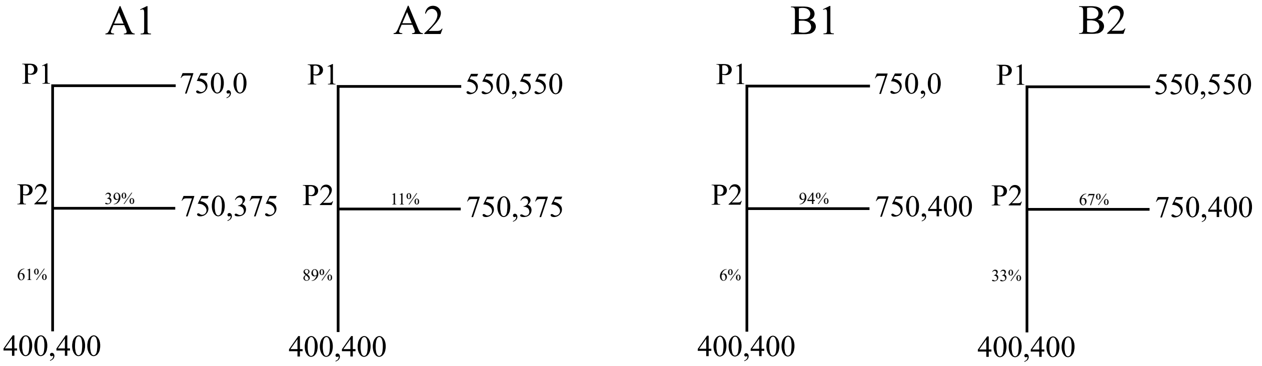

Figure 3: The games analyzed in Charness and Rabin (2002). In this study A1 is coded Berk21; A2 is combined Barc1 and Berk13; B1 is Barc7; and B2 is Barc5.

From the perspective of our theory, assuming log utility of money, games A1/B1 are very different from games A2/B2 because (550, 550) is the most appropriate payoff in the latter, whereas (750, 0) is the least appropriate payoff in the former (again renormalizing to allow comparison). Thus, in the A2/B2 games P2 should punish P1 for not choosing the most appropriate outcome, but in A1/B1 no norm is violated when P1 continues the game and so no punishment is expected.[11]

When P1 chooses to pass the decision to P2 in A2/B2, P2 resents P1 for doing so, and the punishment norm is activated. This changes the normative evaluation of the remaining actions, such that it becomes appropriate for P2 to minimize P1's payoff by choosing (400, 400). To the extent that subjects care about following the norm, they should be more likely to choose this allocation in A2/B2 than in A1/B1. This is exactly what CR report: in the A games the proportion of P2s who choose (400, 400) increases from 61% to 89%, and in the B games from 6% to 33%. Notice that overall more P2s in the A games choose (400, 400) than in the B games because of the material incentive to gain 25 points, which is absent in the B games. Moreover, the outcomes (750, 375) and (750, 400) in the A and B games respectively have higher appropriateness than

[11] The changes in the normative valences of the outcomes (400, 400) and (750, 375/400) due to the change in the third outcome ((750, 0) vs. (550, 550)) are minimal and do not play much role in our reasoning.

the outcome $(400, 400)$, which is consistent with the observation that a non-negligible number of subjects choose these options. □

Case 1 shows that our model can account for the comparative statics of punishment rates in simple extensive form games. In second- and third-party punishment games with an external punishment mechanism, we can test our theory of punishment more directly. In the norm-dependent preferences framework, third-party punishment is not a particularly surprising phenomenon. Since punishment of norm violators is also a norm, anyone with high enough propensity to follow norms, including third parties, should be willing to pay to punish a violator. The fact that many studies report costly punishment by third parties supports this idea (Fehr and Fischbacher, 2004; Leibbrandt and López-Pérez, 2012; Balafoutas et al., 2014; Nikiforakis and Mitchell, 2014). We analyze the seminal study by Fehr and Fischbacher (2004).

Case 2. Fehr and Fischbacher (2004). In the experiment by FF, subjects play the standard DG. However, after the game, third and second parties can punish the dictator, paying 1 unit of personal cost to impose 3 units of cost on the dictator. Subjects choose punishment levels via the strategy method for all possible offers that could be made by the dictator.

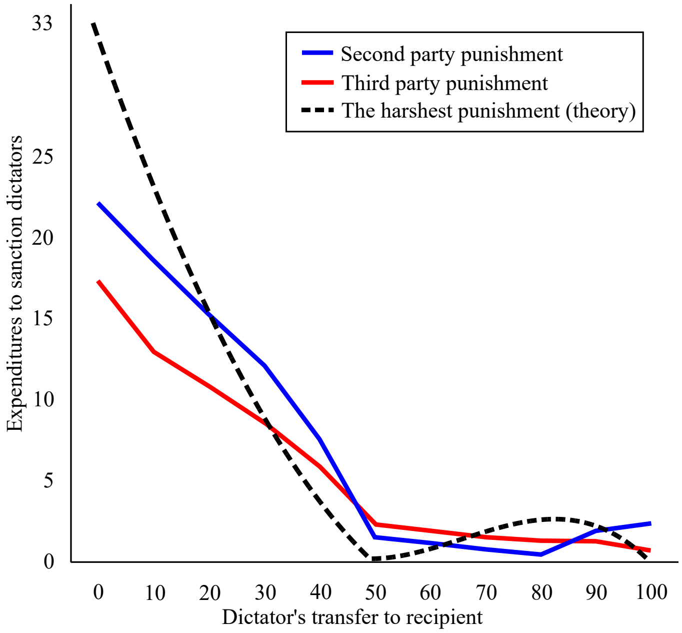

Figure 4: Third and second party punishment in the DG reported in FF. The dashed line shows the model predictions of the harshest possible punishment meted by the extreme norm-followers with very little costs of punishment.

Figure 4 shows the observed levels of punishment by third and second parties alongside the predictions of our model when costs of punishment are negligible and rule-following propensity is very high. Thus, the dashed line plots the upper bound on the amount of punishment that our theory predicts. In accordance with what we called the proportionality principle, the amount of punishment observed in the experiment grows with the distance from the equal split, whenever the dictator gives less than half of the pie to the recipient. Moreover, negligible punishment observed in the cases in which the dictator gives more than half of the pie is also consistent with proportionality, since maximum punishment only involves reducing the violator's payoff to the level of her minimum possible payoff in the game, which is zero in the DG.

The data are also consistent with our deterrence principle: they show that the average punishment strategy reported by subjects makes it unprofitable to give less than half to the recipient (Figure 6 in FF). Our model predicts a strikingly similar pattern of punishment. The fact that the observed punishment is less than the harshest punishment predicted in our model is not surprising, since not all subjects have high propensity to follow norms: ϕ_i also influences the willingness to punish norm violations. Thus, subjects with low/intermediate ϕ_i may prefer to avoid the costs of punishment because they value money.[12] □

Finally, in Case 3 we analyze the third party punishment behavior in Prisoner's Dilemma reported by Fehr and Fischbacher (2004). Our model suggests an explanation of the puzzling observation that subjects punish defectors less after outcome (Defect, Defect) than after outcome (Defect, Cooperate).

Case 3. Prisoner's Dilemma in Fehr and Fischbacher (2004). FF also analyze third party punishment of participants in a Prisoner's Dilemma (PD). They find that cooperation by both subjects is not punished (expenditure of around 0.07, not significantly different from zero); defection by subjects paired with a cooperator is punished the most (expenditure 3.35); and defection by subjects paired with another defector are punished somewhat, but less extensively (expenditure 0.58). Intuitively, applying our model to the PD suggests that players who cooperate should never be punished, since this choice is always consistent with trying to achieve the normatively best outcome. This is consistent with the data. However, whether defection should be punished depends on the payoff parameters of the PD and the players' norm-following propensities: the model predicts that defection should always be punished in a PD with parameters under which norm-dependent utility transforms the game into one with a unique cooperative Nash equilibrium; under such conditions, defection is a clear norm violation. However, when norm-dependent utility merely transforms the PD into a coordination game (see Kimbrough and Vostroknutov, 2023, Example 2), it can be appropriate for even a norm-follower to defect if they believe that others will defect too. So, in this case the justification for punishment becomes less clear. Given this uncertainty about the game, third party punishers may interpret defection in the outcome cooperate-defect as a signal that the defector is violating the norm (which implies high punishment), but defection in the outcome defect-defect as a possible strategic play of two norm-followers (less punishment). Thus, our model also helps to organize these observations from FF that are hard to interpret with other models. □

[12] It is also possible that norms of punishment specify less-than-complete retribution as captured by the parameter σ in our model.

5 Conclusion

While norm violations are not the only motive for punitive action, norm-driven punishment plays an especially important societal role because it helps sustain cooperative and pro-social norms. We argue that a model of endogenous norms developed by Kimbrough and Vostroknutov (2023) also contains within it a plausible account of norm-driven punishment. The model defines the normatively best outcome as the one that minimizes the "aggregate dissatisfaction" across all interested parties in a game or choice set; norm violations, then, can be defined as actions that make it impossible to achieve the best outcome. Moreover, the degree of violation can also be endogenously defined in terms of the gap between the normatively best outcome overall and the normatively best outcome that remains feasible after the violation occurs. We build a model in which this degree of norm violation is proportional to the *resentment* that observers feel and hence to the likelihood and magnitude of punishment.

We argue that punishment of violations can be understood as a meta-norm that exists to sustain other norms, and we show how to embed this norm into game theoretic models with players who have norm-dependent utility. Such a model can be analyzed with standard tools and makes predictions about behavior and punishment. We test the predictions about punishment by collating evidence from previous studies of second- and third-party punishment in games by Charness and Rabin (2002) and Fehr and Fischbacher (2004), and we show that our model accounts for observed differences in punishment across settings and across decisions within a setting. More broadly, there also is evidence corroborating the view that *effective* punishment is norm-driven. In particular, Xiao (2018); Bicchieri et al. (2021) both show that punishment is most effective when subjects are informed that it results from a norm violation.

The notion that normative judgments and punishment arise from emotional reactions to realized and forgone outcomes has a long history (Hume, 1740; Smith, 1759; Prinz, 2007; Smith and Wilson, 2019). The idea is that emotions are stirred when we compare the world as it is to the world as we think it ought to be (or wish it were). Among children, these emotional reactions are mostly driven by self-interest and result in costly "punishment", e.g. children throw tantrums when a parent refuses to indulge them. However, as children mature, the capacity for empathy allows them to recognize similar sentiments in others. Norms emerge such that their view about how the world ought to be begins to account not only for their own interests but also the interests of others. Eventually people develop the capacity to evaluate their own actions and those of others from the abstract point of view of shared norms (Tomasello, 2019). Reactions to the violation of these norms also come to motivate punishment; see also Akerlof (2016) who builds a model in which anger about violations motivates the enforcement of rules.

That punishment has an emotional component has also been demonstrated in laboratory studies of the ultimatum game. For example, Pillutla and Murnighan (1996) attribute rejections of unequal offers to expressions of anger. Consistent with this view, Xiao and Houser (2005) show that allowing opportunities for alternative modes of emotional expression can actually reduce rejection rates, and Grimm and Mengel (2011) show that giving responders a mandatory "cooling off" period also reduces rejections. However, these experiments do not distinguish punishment that arises from anger about not getting the best outcome (from the point of view of ego) from punishment that arises from a failure to achieve the best outcome (as defined from the point of view of a norm). For this purpose, third-party punishment games are more suitable because they remove the impact of second-party anger and thus provide more direct evidence of norm-driven resentment. To be clear - we do not doubt that the motivation for punishment is richer than "mere" concern about norm violation; our model simply shows how to define norm-driven punishment in terms of endogenous norms and then shows that patterns of observed punishment behavior are consistent with the model. A fuller account in which both retributive and norm-enforcing motivations operate side by side is a fruitful direction for future work. Distinguish these motivations may help us understand when and why punishment is likely to result in "backfire" or counter-punishment (Xiao, 2018; Balafoutas et al., 2014).

References

Akerlof, R. (2016). Anger and enforcement. *Journal of Economic Behavior & Organization*, 126:110–124.

Balafoutas, L., Grechenig, K., and Nikiforakis, N. (2014). Third-party punishment and counter-punishment in one-shot interactions. *Economics letters*, 122(2):308–310.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C., Dimant, E., and Xiao, E. (2021). Deviant or wrong? The effects of norm information on the efficacy of punishment. *Journal of Economic Behavior & Organization*, 188:209–235.

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., and Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review*, 97(3):818–827.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117:817–869.

Chudek, M. and Henrich, J. (2011). Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends in cognitive sciences*, 15(5):218–226.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47:268–298.

Elster, J. (1989). Social norms and economic theory. *The Journal of Economic Perspectives*, 3(4):99–117.

Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54:293–315.

Fehr, E. and Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and human behavior*, 25(2):63–87.

Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994.

Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868):137.

Grimm, V. and Mengel, F. (2011). Let me sleep on it: Delay reduces rejection rates in ultimatum games. *Economics Letters*, 111(2):113–115.

Güth, W. and Kocher, M. G. (2014). More than thirty years of ultimatum bargaining experiments: Motives, variations, and a survey of the recent literature. *Journal of Economic Behavior & Organization*, 108:396–409.

Güth, W., Schmittberger, R., and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, 3(4):367–388.

Henrich, J. (2015). *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Hume, D. (1740). *A Treatise of Human Nature*. Oxford: Oxford University Press, (2003) edition.

Kandori, M. (1992). Social norms and community enforcement. *Review of Economic Studies*, 59(1):63–80.

Kessler, J. B. and Leider, S. (2012). Norms and contracting. *Management Science*, 58(1):62–77.

Kimbrough, E. and Vostroknutov, A. (2016). Norms make preferences social. *Journal of European Economic Association*, 14(3):608–638.

Kimbrough, E. and Vostroknutov, A. (2018). A portable method of eliciting respect for social norms. *Economics Letters*, 168:147–150.

Kimbrough, E. and Vostroknutov, A. (2023). A theory of injunctive norms. mimeo, Chapman University and Maastricht University.

Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: why does dictator game sharing vary? *Journal of European Economic Association*, 11(3):495–524.

Leibbrandt, A. and López-Pérez, R. (2012). An exploration of third and second party punishment in ten simple games. *Journal of Economic Behavior & Organization*, 84(3):753–766.

López-Pérez, R. (2008). Aversion to norm-breaking: A model. *Games and Economic behavior*, 64(1):237–267.

Merguei, N., Strobel, M., and Vostroknutov, A. (2022). Moral opportunism as a consequence of decision making under uncertainty. *Journal of Economic Behavior and Organization*, 197:624–642.

Nikiforakis, N. and Mitchell, H. (2014). Mixing the carrots with the sticks: Third party punishment and reward. *Experimental Economics*, 17(1):1–23.

Osborne, M. J. and Rubinstein, A. (1994). *A course in game theory*. Cambridge, Mass.: MIT Press.

Pillutla, M. M. and Murnighan, J. (1996). Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational Behavior and Human Decision Processes*, 68(3):208–224.

Prinz, J. (2007). *The emotional construction of morals*. Oxford University Press.

Smith, A. (1759). *The Theory of Moral Sentiments*. Liberty Fund: Indianapolis (1982).

Smith, V. L. and Wilson, B. J. (2019). *Humanomics: Moral sentiments and the wealth of nations for the twenty-first century*. Cambridge University Press.

Tate, C., Kumar, R., Murray, J. M., Sanchez-Franco, S., Sarmiento, O. L., Montgomery, S. C., Zhou, H., Ramalingam, A., Krupka, E., Kimbrough, E., et al. (2022). The personality and cognitive traits associated with adolescents' sensitivity to social norms. *Scientific Reports*, 12(1):15247.

Tomasello, M. (2019). *Becoming human: A theory of ontogeny*. Harvard University Press.

Xiao, E. (2018). Punishment, social norms, and cooperation. In *Research Handbook on Behavioral Law and Economics*, pages 155–173. Edward Elgar Publishing.

Xiao, E. and Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 102(20):7398–7401.