A Theory of Injunctive Norms[*]

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

April 2021

Abstract

Theories of norm-dependent utility assume commonly known injunctive norms that rank feasible outcomes by their normative valence, but as yet normative valences have only been measured experimentally. We provide a theoretical foundation that assigns a normative valence to each outcome based on players' dissatisfactions, which depend on the higher utilities that they could have received at other outcomes. The normatively best outcome is the one that minimizes aggregated dissatisfaction. Our model imposes structure on theories of norm-driven behavior, rendering them precise and falsifiable. We consider a variety of illustrative applications, highlighting the intuition and explanatory power of the model.

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

[*] We would like to thank Valerio Capraro, Hartmut Kliemt, Matthias Stefan 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, and seminars at GSBE-ETBC Maastricht, George Mason, UC-Davis and Virginia Tech 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

Decades of experimental studies of games played by strangers have revealed that people regularly help, cooperate, share, trust, reciprocate, contribute, reward, and punish, even when doing so is inconsistent with material payoff maximization. To account for these observations in a unifying framework, economists have proposed that such prosocial behavior reflects an intrinsic desire to adhere to commonly known (injunctive) social norms (Cappelen et al., 2007; López-Pérez, 2008; Kessler and Leider, 2012; Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2016); such norms are meant to capture shared agreement about the social appropriateness (and inappropriateness) of various outcomes. Thus, these models assume that people consider not only what they *want* to do (from the point of view of payoff maximization), but also what they *ought* to do from the point of view of shared injunctive norms. When norm-adherence and payoff-maximization conflict, people face a tradeoff, which is sometimes resolved in favor of following the norm. It is straightforward to formulate this intuition in a utility function that associates each outcome with both a payoff and a normative valence and thereby to generate predictions about the influence of norms on behavior.

These models have been shown to have considerable explanatory power, but a lingering concern has been that they provide the researcher with too many "modeler degrees of freedom." To give an analogy, payoff-maximization was not adopted as an axiom for its psychological plausibility; rather it seems to be a straightforward application of the approach championed by Stigler and Becker (1977), who suggest that economists ought to tie their hands by committing to the view that people have broadly similar, consistent and stable preferences and restricting potential explanations for behavioral differences to differences in income and prices. Similarly, defenders of the social norms approach can point to explanatory successes (e.g., the ability to account for observed effects of supposedly irrelevant changes to the action set, as in Krupka and Weber, 2013), but critics (rightly) note that, without a theoretical account of *how* norms vary across choice settings, there is little to discipline the set of possible predictions. Thus far, the response to these concerns has been to use elicitation techniques to measure shared normative beliefs directly in each setting, allowing subjects' reports on social norms to constrain the theory (see e.g., Kimbrough and Vostroknutov, 2018; Chang et al., 2019). With a little ingenuity, techniques like those introduced in Krupka and Weber (2013) are readily adapted to elicit norms in any context, but such an approach still implicitly allows norms to vary arbitrarily across subject pools and choice settings. These concerns call for a theory that imposes some structure on norms.

With these matters in mind, we propose a simple, tractable, and falsifiable theory of injunctive norms in games and allocation problems that can be applied using the standard tools of game theory. We assume that agents' utilities are well-described by "norm-dependent preferences" of the kind mentioned above, and we present a theory that offers an account of the *normative valence* of available outcomes (or consequences). That is, we take for granted that norm-following is a reasonable model of behavior, and we provide a theory of the *content* of the norms by which individuals' decisions are shaped under such a model. With the injunctive norm pinned down, one can simply plug this piece into the norm-dependent utility function to make predictions.

Our approach evaluates the normative valence of each possible consequence in the context of all other possible consequences. The idea is straightforward: the appropriateness of the consequence that I choose (or that we achieve) depends on the set of other consequences that I might have chosen (or that we might have achieved) instead. We ground the evaluation of consequences in the psychology of dissatisfaction: each consequence has a utility associated with it, and consequences worse for me than others that might have attained evoke dissatisfaction.[1] That is, we assume that each agent's fundamentally self-interested desire to achieve better outcomes for himself generates dissatisfaction when less preferred outcomes are achieved. We then assume that normative agreement is founded on common acknowledgment of this source of dissatisfaction, and thus, to define the normative valence of a particular outcome, we aggregate dissatisfaction across all interested parties. The most socially appropriate consequence is simply the one that minimizes aggregate dissatisfaction aggregated across individuals, and the least socially appropriate consequence is the one that maximizes aggregate dissatisfaction.[2] Agents are other-regarding not directly, in the sense of caring about others' utility, but indirectly, insofar as norms account for the dissatisfaction of all.

Our approach to generate a normative ranking of outcomes is not totally new – the notion of Pareto optimality of an outcome, at the core of neoclassical welfare economics, also considers the set of all possible outcomes and aggregates across individuals in a similar fashion. Indeed, there is considerable overlap between our approach and the standard approach in the sense that all Pareto improvements are considered normative improvements under our theory. However, our theory goes a step further in (almost always) providing criteria for choosing from among a set of Pareto optimal allocations. In our model, not all Pareto optimal allocations generate the same aggregate dissatisfaction.

Moreover, the idea that consideration of our own and others' dissatisfaction can provide constraints on what constitutes normatively acceptable behavior is also not new. Here we draw on a rich tradition that grounds the moral sense in our emotional responses to both attained and foregone outcomes and in our ability to empathize with others, understanding the emotions that they might feel in similar circumstances (Hume, 1740; Smith, 1759; Mackie, 1982; Prinz, 2007). Like Adam Smith, we start with the assumption that people are motivated by their own interests; they prefer certain outcomes and resent actions taken by others to prevent those outcomes from being achieved. Yet we also assume that individuals consider the consequences of their behavior

[1] There is a clear connection to the idea of regret, but in the model when norms are defined, this dissatisfaction is prospective rather than retrospective. An appropriate term might be "pre-gret."

[2] In a companion paper (Kimbrough and Vostroknutov, 2021) we develop an axiomatic foundation for this construction.

for others, with "fellow-feeling" allowing us to more-or-less understand how others might feel should a particular consequence attain and hence with normative judgments calibrated to temper naked self-interest, bringing actions into line with what others will "go along with" (Smith and Wilson, 2017).[3] The key to our theory is the implication that what others will go along with depends on what other possible outcomes are available to them.

To our knowledge, ours is the first model attempting to account for the *content* of injunctive norms in laboratory environments. As such, we certainly do not view this as the last word on the subject. Instead, our goal is to show one way the problem can be approached, and then we take the model to data. In particular, we assess the interpretive and predictive power of the model against experimental evidence collated from a diverse set of studies from the literature on prosocial behavior and social preferences.

Experiments have identified situations in which behavior is neatly characterized as inequity averse, maximin, efficiency-seeking, reciprocal, punitive, and so on; this evidence has been summarized and interpreted through the lens of social preferences, with new utility specifications introduced to account for new observations. As noted above, economists working in the social norms paradigm have pointed out that such diversity of behavior across settings can be reconciled if one invokes a preference for following norms and allows norms to vary across settings. Rhetorically, such accounts have attempted to be ecumenical, understanding particular social preference models as special cases of the social norms model, with a particular social preference formulation reflecting a particular environment-specific social norm. We also take this view, and as we show below, behaviors predicted by each of the aforementioned kinds of social preferences also arise naturally from our theory in particular environments (see also Kimbrough and Vostroknutov, 2021, for more discussion of this connection).

We consider three applications of our theory, identifying norms and working out the implications for behavior in 1) settings where subjects make choices over a set of simple two- or three-person resource allocation vectors, 2) games that differ from one another only by the addition or subtraction of possible outcomes, and 3) second- and third-party punishment games. We focus on these examples because each highlights an important feature of the theory and because in each case, the theory makes clear predictions that we can assess in light of existing experimental evidence. In our Appendix we consider a variety of additional applications for the curious reader.[4]

We study allocation problems because these allow us to highlight precisely how norms are shaped by the set of possible outcomes under the theory. Analyzing simple dictator games we

[3] One interpretation of our theory could be in terms of a contractarian approach to morality, in which moral rules are reached by the mutual consent of those who will abide by them (Sugden, 2018). It is interesting to ask whether one could show that the dissatisfaction-minimizing norms are those to which people would be most likely to mutually consent.

[4] Recently, our model has been used to study certain regularities in normative behavior (Merguei et al., 2020; Panizza et al., 2021). These studies also provide a direct (and rather successful) test of the theory.

show that the theory predicts how agents choose one Pareto optimum among many. Analyzing the games studied by Engelmann and Strobel (2004) and Galeotti et al. (2018), we show how norms vary with the choice set, yielding norms that favor efficiency over equality in some cases, equality over efficiency in others, and maximin if we assume sufficiently concave utility over money. Thus, we highlight how our model connects to the social preferences literature – potentially helping to explain why measured social preferences vary across environments.

We study the second set of games to highlight the implications of our assumption that the normative evaluation of any one outcome depends on the entire set of possible outcomes. This implies that, as the set of outcomes is expanded or contracted, the normative evaluation of the remaining outcomes may change. In this context, we study the modified dictator games of List (2007) and the voluntary and involuntary trust games of McCabe et al. (2003), in which it has been shown that adding (respectively, subtracting) an outcome has notable impact on behavior, and we show that these observations can be interpreted naturally in the context of our model as the dissatisfaction with one outcome is directly affected by the introduction (or removal) of another.

Finally, we study the third set of games to show how our work connects to existing experiments on the emergence and maintenance of norms. It is well-established that norms emerge and are sustained by punishment of violations (Kandori, 1992; Henrich, 2015); we argue that such punishment is driven by *resentment* of actions that violate norms. We analyze second-party punishment in a set of games due to Charness and Rabin (2002) and the third-party punishment games introduced by Fehr and Fischbacher (2004) to highlight how our 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).

For simplicity, we have thus far assumed 1) that each party has an equal (or non-existent) prior claim to the resources being allocated and 2) that norms are defined impartially, such that we treat each outcome and each individual's dissatisfaction equally in aggregation to determine the norm. Such a model captures the normative valence of outcomes in games played among coequal strangers, with no prior claims and minimal environmental cues about what behavior(s) are appropriate or inappropriate. In this sense, *the basic theory is addressed to the spare contexts typically studied in economics experiments.*

Thus, the model we have described captures variation in norms by richly accounting for the context in which each choice is made (in terms of counterfactual outcomes), but it does not account for other aspects of "social context" that are known to influence behavior in the lab (e.g., ownership, role-assignment protocols, in- and out-group). We next show how to extend the model to account for the ample evidence that "context matters" and the implication that norms often depend on such context.

In particular, we take as axiomatic the natural human tendencies to respect ownership and entitlement claims and to favor kin, in-groups and high-status individuals in the moral calculus.

Then, we show that all of these can be handled in a straightforward way by applying appropriate weights to individuals' dissatisfaction during aggregation.

First, we show that ownership claims to some or all of an endowment can be handled by weighting the dissatisfaction associated with a (counterfactual) reduction in payoffs by the strength of each player's ownership claim. Thus, the normative valence of each outcome can be made to depend, in a natural way, on the strength of the prior claim that each player has to the pie. We then show how entitlement to a "role" (e.g., when someone has earned the right to be a decision-maker) can be handled by assuming that such entitlements reduce others' resentment of norm violations; someone who is entitled to a role is therefore less punishment-worthy than someone who has no such entitlement. Finally, we show that the model can account for norms of differential treatment of in- and out-groups, high status people, and kin if, when aggregating dissatisfaction across individuals to define the norm function, we weight the dissatisfaction of others in proportion to the degree of kinship or status. To properly use the model and retain falsifiability, it is essential that these weights be known (or estimated) prior to and separately from the environment being studied. Thus, we illustrate these intuitions with examples from the literature in which the weights can be estimated from prior data (e.g., via a within-subject experimental design) or drawn from theory (e.g., coefficients of kinship from biological theory) and then used to make predictions.

At this point, it is worthwhile to highlight what our model does not do: first, it is not intended as a one-size-fits-all explanation of all norms. Nothing about our theory precludes the existence of other "norms" defined in the sense often used by game theorists. That is, we have no doubt that regularities of social behavior often arise as equilibria of (repeated) games, and many such norms may be Pareto suboptimal. Our model is supposed to capture *injunctive* norms; in our view, these are an input into the game theoretic analysis that determines actual patterns of behavior, but they don't determine social behavior all by themselves. Strategic considerations remain relevant. That said, we think our model can help to understand the standpoint from which people criticize extant suboptimal norms: by combining counterfactual comparisons and empathy.[5] Second, we make a number of simplifying assumptions for ease of exposition that are not likely to hold in practice. For instance, we assume common knowledge of (and agreement on) whose dissatisfaction "counts" and how much in defining what is appropriate. However, we see the fact that this assumption may be violated in practice as instructive: in our view, many cases of normative disagreement stem from disagreement over who (or what) should count, and how much.[6] Similarly, we assume an implausibly powerful ability to empathize with others, requiring complete knowledge of others' utility functions to get norms off the ground. This highlights another important source of normative disagreement: lacking knowledge of others' preferences

[5] For example, when we criticize norms of female genital mutilation or child marriage, we do so by considering how much better off the victims could be in other circumstances.

[6] For example, whether a non-vegan diet is normatively appropriate depends on whether (and how much) we count animals in our normative calculus.

and choice sets, we often incorrectly judge the actions of others and fail in our attempts to do good (so that our good intentions are thereby misinterpreted). We hope that in mapping out the boundaries of our theory we highlight the right kinds of issues to improve our understanding of social behavior. If a reader has come with us this far, then we hope that these complications will be a wellspring of future research.

In sum, we present a theory of injunctive norms meant to complement existing work in both the social norms and social preferences frameworks. The theory grounds norms in the psychology of dissatisfaction. Dissatisfaction with a particular outcome is defined relative to *all* other feasible outcomes, such that the evaluation of any particular outcome depends on what other outcomes are possible. We assume that norms reflect the aggregation of such prospective dissatisfaction across individuals; that is, normative judgments arise from considering how (dis)satisfied the self and others would be with a particular outcome (relative to alternatives), with the normatively best outcome being the one that minimizes the aggregated dissatisfaction of interested parties. We work out the implications of the model and confirm its interpretive power by identifying how norms vary across a variety of experimental games and showing that behavior in variants of those games changes in a manner consistent with changes in norms.

2 Model

2.1 Definition of Normative Valence

We begin with a definition of normative valence. Intuitively, this notion is meant to capture shared beliefs about the appropriateness of an outcome. In what follows, we assume that normative valences depend on the final outcomes of a game and not on its strategic structure defined by a sequence of moves, information sets, etc. Therefore, we start with a set C of *consequences* 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 consequence a vector of players' utilities (payoffs) with $u_i(x)$ meaning the payoff of player i for consequence $x \in C$.[7]

As noted above, we define the normative valence of an outcome in terms of comparative *dissatisfaction*, with the normatively most appropriate consequence in the set of possible consequences, which we will call a *norm*, being the dissatisfaction-minimizing consequence. Thus, we start with our definition of dissatisfaction for a particular consequence, and then we explain how we aggregate across (counterfactual) consequences within an individual, and finally how

[7] The definition of normative valence below is based on the payoffs defined by the function u. Thus, instead of having a separate set of consequences and a utility function, we could have assumed that consequences *are* the payoff vectors. However, this would not allow us to distinguish cases in which several consequences result in the same payoff. This distinction turns out to generate important (testable) implications. See Example 4 in Appendix A for details and Panizza et al. (2021) where the influence of "repeated" consequences is tested directly.

we aggregate dissatisfaction across individuals to define the normative valence of each consequence.

The main ingredient of our definition of normative valence is

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

the *dissatisfaction* that player i feels about consequence x *because of* the possibility of c. This notion of dissatisfaction is intended to capture attention to foregone possibilities. Thus, we assume that if consequence x attains, then player i suffers dissatisfaction from it to an extent $d_i(x, c)$ because c could have attained instead. Dissatisfaction is positive when c brings player i more utility than x and zero otherwise.[8]

Next we define the aggregation of dissatisfaction within an individual, which we assume depends on the entire set of possible counterfactual consequences. Let

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

denote the *personal dissatisfaction* that player i feels with respect to x. Thus, we assume that a low utility outcome results in more (less) dissatisfaction the larger (smaller) is the set of counterfactual higher-utility outcomes. Intuitively, this reflects the idea that one's view of their present circumstances may deteriorate upon the emergence of new opportunities that might make them better off.[9]

Next, we define the *aggregate dissatisfaction* of x—that is, dissatisfaction aggregated across all individuals—as

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

The function D captures the dissatisfaction of all players for each possible consequence in a game. This second form of aggregation reflects our assumption that aggregate dissatisfaction of the players depends only on personal dissatisfactions.[10] This aggregation is intended to capture empathy, with individuals applying their knowledge of how others would feel at a given outcome to agree upon a normative ranking.

[8] In a companion paper (Kimbrough and Vostroknutov, 2021), we provide an axiomatic foundation for our theory. There we show how this particular shape for the dissatisfaction function follows from two axioms. That we should take the difference in utilities is implied by the assumption that dissatisfaction does not depend on the overall level of utility (i.e., adding a constant to the utilities of a player in all consequences does not change any player's dissatisfaction). The max operator reflects the assumption that one's evaluation of x does not improve simply by adding a less preferred option to the set. This implies that adding Pareto dominated consequences does not change dissatisfaction.

[9] In Kimbrough and Vostroknutov (2021) this is stated as an axiom: for any set of consequences C that includes x, adding another consequence that gives i higher utility than $u_i(x)$ makes her feel more dissatisfaction from x.

[10] Axiomatically, aggregate dissatisfaction of x is constant as long as all personal dissatisfactions are constant, regardless of which particular consequences cause each player to be dissatisfied with x personally (Kimbrough and Vostroknutov, 2021).

Finally, we assume that the normative valence associated with a consequence x is inversely proportional to its aggregate dissatisfaction.[11] Thus, the consequence which generates the least aggregated dissatisfaction is considered the most socially appropriate (the norm), and the consequence with the highest aggregate dissatisfaction the least socially appropriate. This conceptual connection is grounded in the philosophical doctrines mentioned in the introduction (Hume, 1740; Smith, 1759; Mackie, 1982; Prinz, 2007) that trace the roots of morality to the negative emotions that arise from personal circumstances and from our capacity to consider how others might feel in similar circumstances. To put it formally, let $\text{Conv}(-D)$ denote the convex hull of the image of the function $-D(x)$ in $\mathbb{R}$; call $\langle N, C, u, D\rangle$ an *environment*; and consider the following definition:

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

$$\eta_C(x) := [-D(x)]_{\text{Conv}(-D)},$$

where $[\cdot]_{\text{Conv}(-D)}$ is the linear normalization from interval $\text{Conv}(-D)$ to $[-1, 1]$, a ***norm function*** *associated with $\langle N, C, u, D\rangle$. If D is a constant function, set $\eta_C(x) = 1$ for all $x \in C$.*

In this definition, η_C is simply the negative of aggregate dissatisfaction, normalized to the interval $[-1, 1]$. Thus, the consequence x with $\eta_C(x) = 1$ is the most socially appropriate (the norm) and the one with $\eta_C(x) = -1$, the least socially appropriate. If all consequences have the same aggregate dissatisfaction, then we assume that $\eta_C(x) \equiv 1$ for all consequences. This last assumption is important since it guarantees that a most appropriate consequence always exists, which is necessary for the relative comparisons of norms across settings (see discussion in Appendix C).[12]

We continue with several examples that illuminate the properties of η_C, and then we briefly consider how our method compares to a plausible alternative dissatisfaction aggregation method.

Example 1. Payoff Efficiency. It is clear from the definition of d_i above that, all else equal, an increase in utility of a consequence for one player implies a weak increase in its appropriateness, since dissatisfaction of this consequence must weakly decrease. For any choice from a set of two

[11] This feature distinguishes our approach from models of regret aversion (Loomes and Sugden, 1982): our computation of dissatisfaction can be seen as reflecting a form of regret aversion, but where our model differs is that the normative evaluation of a consequence depends on the *aggregation* of this regret across all interested parties. Each agent's choice thus (weakly) depends not only on their own regret but also on the regret of others.

[12] The interval $[-1, 1]$ was chosen arbitrarily. When the analysis is focused on only one choice set C, the exact interval for normalization is irrelevant. However, it becomes very important when different norm functions are compared (as in Appendix C, for example). In this case, the normalization interval for different norms emphasizes how important these norms are relatively to each other. For example, 50 years ago people knew that plastic is bad for the environment, but the norm for recycling it was very weak in comparison to other norms. Today, this norm has grown considerably in relative strength. Panizza et al. (2020) show some evidence pointing towards changing normalization intervals when information about relative (un)importance of the norm arrives.

consequences $C = \{c_1, c_2\}$, it turns out that the consequence with the highest sum of utilities across the N players (highest payoff efficiency) is always the norm. Suppose the utilities of the players for the consequences c_1 and c_2 are given by $(a_1, ..., a_N)$ and $(b_1, ..., b_N)$ and suppose that $a_1 + ... + a_N > b_1 + ... + b_N$, or the efficiency of c_1 is higher than the efficiency of c_2. This inequality can be rewritten as

$$\sum_{i:a_i>b_i} a_i - b_i > \sum_{i:a_i<b_i} b_i - a_i \;\Leftrightarrow\; D(c_2) > D(c_1) \;\Rightarrow\; \eta_C(c_1) = 1 \text{ and } \eta_C(c_2) = -1.$$

Thus, in any set of just two consequences, the more appropriate consequence is the one with the highest payoff efficiency. Notice that this property does not hold anymore if there are more than two consequences. With three consequences, the most payoff efficient one does not necessarily minimize dissatisfaction. This has implications for the measurement of social preferences in allocation decisions to which we return below. □

Example 1 shows that the overall sum of utilities matters for the appropriateness of the consequences, and in case of two consequences this also implies that the payoff efficient one is more appropriate even if the consequences are not Pareto comparable. More generally, it is true that if consequence c_1 Pareto dominates c_2 then c_1 is (weakly) more appropriate than c_2. We formulate this statement as a proposition.

Proposition 1. *In an environment $\langle N, C, u, D\rangle$ if consequence $c_1 \in C$ Pareto dominates $c_2 \in C$ then $\eta_C(c_1) \geq \eta_C(c_2)$ with strict inequality if C is finite.*

Proof. See Appendix E.

Thus, our definition of a norm function respects the core tenet of neoclassical welfare economics – that economic forces should push society towards Pareto optimal outcomes. Nevertheless, as important as the idea of Pareto optimality is for economics, it fails to provide any guidance on how an allocation ought to be chosen on the Pareto frontier. Our notion of an injunctive norm goes further and (usually) provides a criterion for choosing among Pareto optimal outcomes. This is demonstrated by the next example in which all consequences have the same payoff efficiency and are on the Pareto frontier.

Example 2. Dictator Game (DG). Suppose a dictator p has a pie of size 1 and chooses to give $c \leq 1$ to a receiver r (and is left with $1 - c$). The set of consequences is $C = [0, 1]$, and the utilities are given by $u(c) = (u_p(c), u_r(c)) = (1 - c, c)$. For any consequence $c \in C$ the personal dissatisfaction of the dictator is $D_p(c) = c^2/2$, and the personal dissatisfaction of the receiver is $D_r(c) = (1 - c)^2/2$. Thus, aggregate dissatisfaction is given by

$$D(c) = D_p(c) + D_r(c) = \frac{c^2}{2} + \frac{(1-c)^2}{2}.$$

This is an upward sloping parabola which is minimized at $c^* = \frac{1}{2}$. Thus, the norm function η_C is a downward sloping parabola with the equal split being the most socially appropriate consequence and the consequences $c = 0$ and $c = 1$ the least socially appropriate ones. This example demonstrates how a norm favoring equality can emerge from the basically selfish desire of all agents to receive higher payoffs coupled with a regard for the dissatisfactions of others.

To see how dissatisfaction affects the norm when utilities of players are asymmetric, let us assume that the receiver has a different "need" for the pie than the dictator. Suppose the receiver's utility is $u_r(c) = \gamma c$, where $\gamma > 0$. To illustrate, suppose that receiver is in dire circumstances and his γ is very large. Intuition suggests that in this case, it is appropriate to give him more than half. Indeed, if we repeat the calculations above with γ included, we find that the norm is now $c^* = \gamma/(1+\gamma)$, which goes to 1 as γ grows to infinity. So, the model implies that it is socially appropriate to give the receiver larger portions of the pie when she needs it more than the dictator. □

Going back to the standard dictator game, it is important to highlight that the equal division emerges as the most appropriate consequence because of the game's symmetry. In general, however, it is not true that "more equal" allocations are more appropriate than "less equal" ones. Rather, in two-player constant-sum games, as in a DG, dissatisfaction is minimized for a "midpoint" consequence, the consequence that has an equal number of better and worse consequences for both players. We prove this result in a proposition.

Proposition 2. *Suppose $\langle N, C, u, D \rangle$ has two players and K consequences $c_1, c_2, ..., c_K$ with utilities $u_1 \leq u_2 \leq ... \leq u_K$ for one player and $a - u_1 \geq a - u_2 \geq ... \geq a - u_K$ for the other ($a, u_1, ..., u_K \in \mathbb{R}$). Then, for any $j = 1..K-1$, $D(c_{j+1}) - D(c_j) = (2j - K)(u_{j+1} - u_j)$. Thus, the midpoint consequences $c_{\frac{K}{2}}$ and $c_{\frac{K}{2}+1}$, if K is even, and $c_{\frac{K}{2}+\frac{1}{2}}$, if K is odd, are the norm.*

Proof. See Appendix E.

Proposition 2 implies that the most appropriate consequence, in case of constant payoff efficiency, is not the one that is the closest to an equal distribution of utility, as most models of social preferences would suggest, but rather the one that is "equal" in terms of the number of other undesirable consequences available: for the most appropriate consequence this number is the same for both players. Thus, consequences which are very unequal in terms of utilities, can still be considered normatively appropriate in specific contexts where most consequences give a large portion of the pie to one player.

Propositions 1 and 2 show that both payoff efficiency and equality play a role in the concept of normative valence that we propose. In general games, however, the two notions can become intertwined in non-trivial ways and it can be difficult to succinctly summarize the implications of the injunctive norm; we consider the problem of how to summarize the injunctive norms calculated under our model more thoroughly in Kimbrough and Vostroknutov (2021).

Our concept of normative valence can also account for other types of social preferences. In some studies (e.g., Engelmann and Strobel, 2004; Baader and Vostroknutov, 2017), which we discuss in more detail below, it is pointed out that many subjects' choices seem to be guided by *maximin preferences* conceptualized by Rawls (1971). In Appendix B we show that maximin preferences can be expressed in our model if we assume diminishing marginal utility of money, which essentially makes maximin a special case of efficiency preferences (see Section 3.1). The general logic of how preferences for maximin emerge is similar to the situation described in Example 2 when players had different utilities of money: with concave utility, "poor" players suffer more dissatisfaction from similar outcomes than their "rich" counterparts. Thus, the norm favors allocating more payoff to poor players.

Finally, we compare our definition of a norm function with two possible alternatives and provide intuitive arguments that support our modeling choices. First, we only consider dissatisfaction with higher counterfactual utilities and not "elation" created by consequences that give less utility. The reason for this is simple: the model with elation often predicts that the asymmetric outcomes cooperate/defect and defect/cooperate are equilibria in the Prisoner's Dilemma, which we find unappealing on evolutionary grounds: agents who follow elation-based norms do not cooperate as much as those who follow dissatisfaction-based norms. We believe that even though humans are definitely capable of feeling such elation (as in the expression "it could have been worse"), they do not use it in moral calculus.

Second, we define personal dissatisfaction of a consequence x as an integral of dissatisfactions of x because of all other consequences in C (equation 2), which follows from the axioms described in Kimbrough and Vostroknutov (2021). One alternative to our way of integration of dissatisfactions is counting only the highest dissatisfaction that each consequence achieves. Mathematically, this can be expressed as $D_i(x) = \max_{c \in C} d_i(x, c)$, which is similar to the formulation proposed in Cox et al. (2018). Let us check the properties of the norm function defined using this personal dissatisfaction formula. Notice that $\max_{c \in C} d_i(x, c) = u_i^* - u_i(x)$, where u_i^* is the highest utility that player i can enjoy. Therefore, $D(x) = u^* - \sum_{i \in N} u_i(x)$, where u^* is the sum of highest payoffs of all players. This means that $\eta_C(x)$ is a positive affine transformation of the sum of payoffs in x. So, the alternative dissatisfaction integration method ranks consequences according to their payoff efficiency and does not differentiate among consequences with a fixed sum of payoffs. In addition to not capturing equality, as our model does, this alternative method of aggregating dissatisfaction ignores the *number* of counterfactual consequences. Our specification does take this into account, which we believe is important. We demonstrate this in Example 4 in Appendix A (also see Panizza et al., 2021). Taken together, these arguments provide support in favor of our model. In Kimbrough and Vostroknutov (2021) we discuss different aggregation methods and their implications for norms in much more detail.

2.2 Punishment

Our model of normative valence is incomplete without a punishment mechanism that would maintain norm compliance by deterring selfish urges to gain more payoff at the expense of breaking the norm. There is a wide consensus that without such punishment mechanisms, the evolution of social norms and preferences for following them would be impossible (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. The authors show experimentally that altruistic punishment is indeed a common phenomenon.[13] From the normative perspective, punishment of violators is a norm in itself and is thus followed in the same way other norms are. 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.

We use this idea to model normative reaction to an action by a player, which is not consistent with achieving the most socially appropriate consequence. 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. 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 secondary norm (of punishment) is activated when the primary norm is not followed in a particular setting.

We start with an environment $\langle N, C, u, D \rangle$ that defines the normative valences of consequences and assume that player i chooses one of the actions from some set A. Each choice $a \in A$ restricts the set of reachable consequences to $C_a \subseteq C$, which nevertheless does not change the norm function η_C associated with them since player i made her choices taking all possible consequences into account. Let $M = \text{argmax}_{c \in C}\, \eta_C(c)$ be the set of the most socially appropriate consequences, 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 consequences unreachable. We consider such a choice a norm violation. The *resentment* of this violation is measured by the difference between the normative valence for the most socially appropriate consequence $\max_{c \in C} \eta_C(c)$, which now cannot be reached, and the most socially appropriate consequence

[13] In our opinion the term "altruistic punishment" is not very well suited for the description of this phenomenon. After all, the incentives to punish come from the desire to follow norms, which does not have to be altruistic; rather, it so happens that many norms favor cooperation, equality and other 'pro-social' outcomes and supporting those norms thus has the flavor of altruism.

that can still be obtained after a was chosen ($\max_{c\in C_a}\eta_C(c)$).[14] Denote this difference by

$$r_a := \max_{c\in C}\eta_C(c) - \max_{c\in C_a}\eta_C(c).$$

The next step is to determine how player i, who violated the norm, should be punished. Punishment is a complex phenomenon, and there might be many reasons for it: the desire to achieve the most appropriate consequence, 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. Therefore, we concentrate on the two core principles that punitive laws are universally based on. One important purpose of punishment is *deterrence*, which means that the amount of punishment should be large enough that a player does not have an incentive to violate the norm. Another is *"an eye for an eye"* principle (EE), which states that the amount of punishment should be proportional to r_a, the degree of norm violation.

To construct the normative valences pertaining to punishment we determine, for each possible payoff of 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 consequence, $u_{im} = \max_{c\in M} u_i(c)$, or the payoff that she chose to forgo when choosing a; 2) the minimal payoff that she can obtain in the game, $\underline{u}_i = \min_{c\in C} u_i(c)$, which serves as a reference point for the harshest punishment possible; and 3) the payoff that i seemingly "aimed at" receiving after choosing a, $\bar{u}_{ia} = \max_{c\in C_a} u_i(c)$.[15] 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 consequence).[16] The EE principle states 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 the 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.

[14] 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 consequence with normative valence $\max_{c\in C_a}\eta_C(c)$ can still be achieved. 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. (2020)). What is important is that the degree of violation is weakly monotonic in r_a (defined below).

[15] 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.

[16] 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 $\bar{u}_{ia}$.

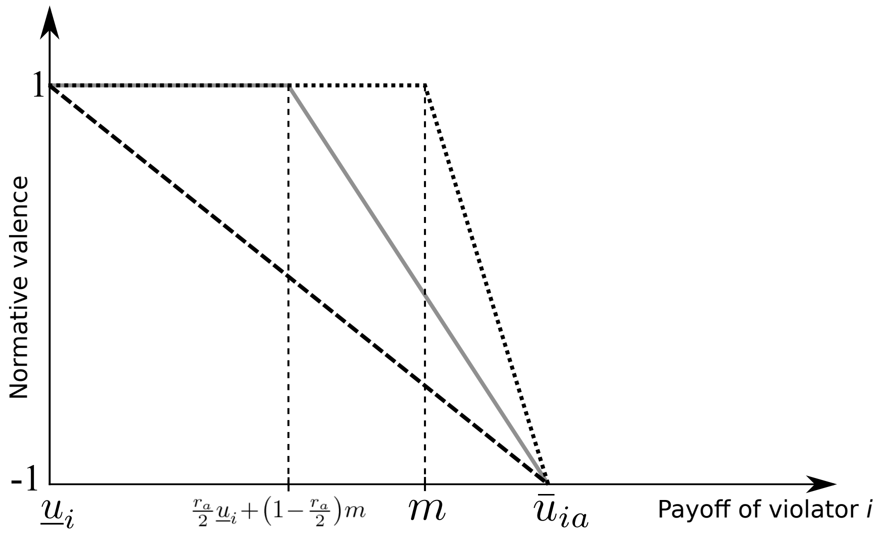


Figure 1: Punishment norms for $r_a < 2$ (solid gray line), $r_a = 2$ (dashed line), and $r_a \to 0$ (dotted line).

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 following. The payoff $\bar{u}_{ia}$, which constitutes i's "criminal 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 $\frac{r_a}{2}\underline{u}_i+\big(1-\frac{r_a}{2}\big)m$. 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 EE 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 $\frac{r_a}{2}\underline{u}_i+\big(1-\frac{r_a}{2}\big)m$ have the highest normative valence of 1. This implies that for the punishers it is normatively irrelevant whether i gets punished by having payoff $\frac{r_a}{2}\underline{u}_i+\big(1-\frac{r_a}{2}\big)m$ or lower.

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 consequence 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 Appendix D.

The second way to implement punishment is by taking retaliatory 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_C and μ_i, thereby, increasing the normative valence of the consequences 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 consequences with $\mu_i(x)$ meaning $\mu_i(u_i(x))$ and assume that for each $x \in C_a$ the combined norm function is

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

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 5 in Appendix A. The intuition is that normatively inappropriate offers by the proposer can "justify" (in the sense of our theory of norms) retaliatory rejection by responders.[17]

2.3 Games with Norm-Dependent Utility

In this section we put all the elements of our model together and analyze how extensive and normal form games with norm-dependent utility are played. Since most games in the experiments that we consider in the next section are rather simple, we restrict the exposition in this section to normal form games and perfect information extensive form games with two moves. The formulation for general games with observable actions can be found in Appendix D.[18]

We start by defining a utility function that takes normative valences as an input. Up to this point our model was purely normative, in the sense that it only described how appropriate or inappropriate the consequences 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

[17] It should be mentioned that Smith (1759) also considered *gratitude* (the opposite of resentment) as an important force that drives human behavior. We can define gratitude similarly to resentment by considering the set of the socially worst consequences $M' = \operatorname{argmin}_{c \in C} \eta_C(c)$ and calculating gratitude for action a with $C_a \cap M' = \emptyset$ as $g_a := \min_{c \in C_a} \eta_C(c) - \min_{c \in C} \eta_C(c)$, or the degree to which action a of player i helped to avoid the worst consequence. The normative reward can be defined using utilities $u'_{im} = \min_{c \in M'} u_i(c)$, the payoff that player i risked receiving by not choosing a; $\underline{u}'_i = \max_{c \in C} u_i(c)$, the reference point for highest reward; and $\bar{u}'_{ia} = \min_{c \in C_a} u_i(c)$, the minimal utility that player i can still get due to his action. The reward can be incorporated into the model through a separate reward mechanism or by combining norms inside the game. We do not consider gratitude further for simplicity, because, unlike punishment, rewards have a potential to lead to Pareto improvements, which we believe should be taken into account when computing norm functions. We leave it for future theoretical and experimental research.

[18] Also, Tremewan and Vostroknutov (2021) show how to model norm transmission and social learning with an extension of this framework that incorporates uncertainty.

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 consequence x as

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

where $u_i(x)$ is the utility of consequence x as defined above, with the set of consequences corresponding to the set of terminal nodes in the game. $\eta(x)$ is the normative valence of x in the game node directly leading to x (pre-terminal node) if no separate punishment mechanism is available.[19] When there exists a punishment mechanism, $\eta(x)$ is the same as $\eta_C(x)$, the norm function defined by all consequences in the game (see Appendix D). $\phi_i \geq 0$ is a constant that defines player i's norm-following propensity (Kimbrough and Vostroknutov, 2016, 2018). 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.[20]

We start with one-shot normal form games. In any such game the set of consequences is the set of outcomes of the game with the payoffs being the consumption utilities defined by the function u. To account for norms we merely need to redefine the game using the norm-dependent utility w instead, and then we can analyze it with standard tools. To illustrate the concept, we analyze the Prisoner's Dilemma in Example 6 in Appendix A.

Next, we analyze extensive form games with two moves and perfect information. Consider a game defined by the environment $\langle N, C, u, D \rangle$ and the observable actions of two players, 1 and 2. First, player 1 chooses an action $a_1 \in A_1$, and then player 2 chooses an action $a_2 \in A_2(a_1)$. Here the set of actions of player 2 depends on the choice of player 1. After the move of player 2 the game ends, and a consequence $c_{a_1 a_2}$ is realized. To understand how the game is played we start with the norm function η_C defined for the tuple $\langle N, C, u, D \rangle$. This function describes the appropriateness of each consequence at history $\{\varnothing\}$, before the game begins.

Next we define the norm function η_{a_1} after player 1 chooses action a_1. Let $C_{a_1} \subseteq C$ denote the set of consequences reachable after a_1 and $M = \text{argmax}_{c \in C}\, \eta(c)$ the set of consequences with the highest appropriateness. For all a_1 such that $C_{a_1} \cap M \neq \varnothing$ or, alternatively, such that game ends after a_1 set $\eta_{a_1} = \eta_C$ (restricted to consequences C_{a_1}). When player 1 chooses an action, which is consistent with eventually achieving some consequence that has the highest appropriateness, the norm function remains unchanged, since this action does not constitute a norm violation (as defined in Section 2.2). Similarly, if the game ends after a_1, there is no need to update the norm function since no punishment is possible. When $C_{a_1} \cap M = \varnothing$ and player 2 has to move, player

[19] As the game unfolds, the norm function can change when an action is taken that merits punishment, so η can be different from the norm function that existed in the beginning of the game. See Appendix D for details.

[20] We think of ϕ_i as a personal characteristic of a player, which is private information. In simple analysis, ϕ_i might be assumed to be common knowledge, but in principle all games should be modeled as those with incomplete information about ϕ_i.

1 has violated the norm and the punishment norm is activated. Thus, the norm function for the remaining consequences C_{a_1} is updated to

$$\eta_{a_1}(c) = \sigma\eta_C(c) + (1-\sigma)\mu_1(c|a_1) \text{ for } c \in C_{a_1},$$

as defined in Section 2.2. Here notation $\mu_1(c|a_1)$ means the punishment norm function that is calculated for the consequences C_{a_1}. Norm functions η_{a_1} are defined on the pre-terminal nodes of the game. Thus, the utility that player i maximizes is given by

$$w_i(c) = u_i(c) + \phi_i\eta_{a_1(c)}(c)\ \forall c \in C.$$

Here $a_1(c)$ is the action of player 1 that leads to consequence c. With the utilities thus redefined one can use any standard game theoretic concept to determine an equilibrium.

3 Evidence

Two recent studies have directly tested the theory presented above: Panizza et al. (2021) show that the model of injunctive norms accommodates behavioral spill-over effects across games and Merguei et al. (2020) use the model of punishment to study moral opportunism. Despite this however, in this section we take the model to the data from a variety of well-known experiments to illustrate how it can be used to interpret behavior in different contexts. First, we show how our model can account for the observation that measured social preferences vary with the task by which they are elicited. Second, we analyze experiments in which a game is expanded or contracted (by adding or removing consequences) and examine how our model accommodates the resulting changes in behavior. Third, we examine the model's implications regarding the actions that constitute norm violations, how these actions ought to be the target of punishment in dynamic interactions, and how severely they ought to be punished. We chose these settings because each of them allows us to illustrate a simple comparative static implication of the theory. Many experimental studies of social interaction are designed in a way that makes interpretation via our model complicated. This is because (under the model) behavior is always influenced by both selfishness and norm-following, and because in many experimental designs, several normative aspects change at once between treatments. For clarity and simplicity, we highlight a set of experiments that change only one normative characteristic of the environment at a time, keeping everything else constant, and which generate comparative static predictions that we can assess with the data.

3.1 Choice-Set-Dependent Social Preferences

In this section we look at a set of experiments from the literature on social preferences, in which subjects make choices among two or three allocations for themselves and others. One of the most well-known papers in this category is Engelmann and Strobel (2004).

Case 1. Engelmann and Strobel (2004). In this experiment subjects choose among allocations for themselves and two other people in multiple tasks (between-subjects). We focus on a subset of these tasks, in which allocations are similar in terms of most payoffs, and we also present data from a replication by Baader and Vostroknutov (2017). Table 1 shows the tasks with three allocations each. The subjects in the role of Person 2 decide which allocation in $\{A, B, C\}$ to implement. Notice that tasks 1, 2, 3 are the same except for the payoffs for Person 2, and similarly are tasks 4, 5. The last two rows show the percentages of subjects who chose each allocation in the two studies. Qualitatively they are very similar.

Task	**1**			**2**			**3**			**4**			**5**		
Allocation	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Person 1	16	13	10	16	13	10	16	13	10	21	17	13	21	17	13
Person 2	8	8	8	9	8	7	7	8	9	9	9	9	12	12	12
Person 3	5	3	1	5	3	1	5	3	1	3	4	5	3	4	5
Choices, %															
ES2004	70	27	3	83	13	3	77	13	10	40	23	37	40	17	43
BV2017	89	8	3	95	4	1	58	14	28	39	14	47	33	14	53

Table 1: Three-person Dictator games that were used in ES and BV.

For each task we calculate two norm functions: the standard one that treats monetary payoffs as utilities (linear utility) and the norm function with log utility. The reason to consider log utility is that the payoff differences in the allocations are rather extreme, Person 3 never gets more than 5 points, while Person 1 gets no less than 10. This can lead to large differences in norm functions between linear and log utility cases (see Appendix B for discussion). Figure 2 presents the norm functions for the five tasks.

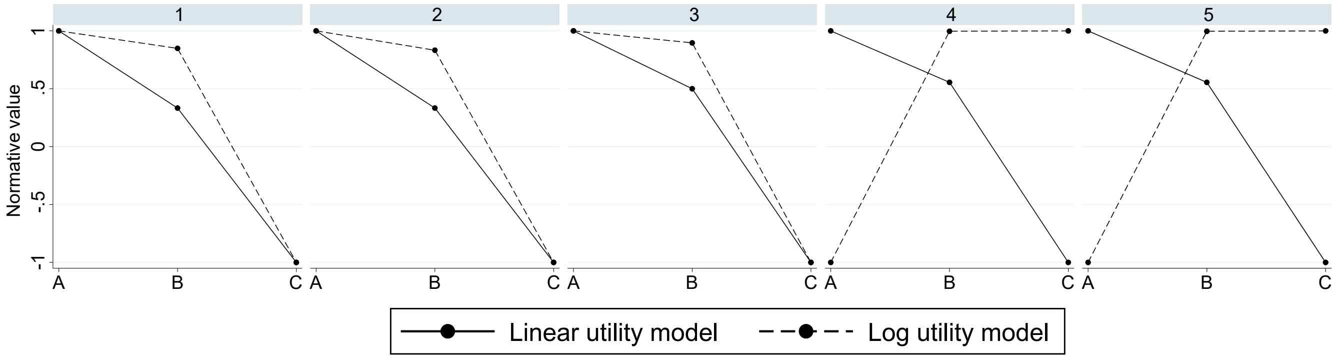


Figure 2: Normative valences in linear and log utility models.

Consider first tasks 1-3. Here the payoffs of Persons 1 and 3 decrease from allocation A to C, which is reflected in their appropriateness. Allocation A is the most appropriate in both linear and log utility models. This is consistent with the subjects' choices: allocation A is preferred by the majority. This is even true for task 3 where Person 2, the decision maker, receives the highest payoff in allocation C. Thus, in tasks 1-3 the norms prescribe the choice of the most efficient allocation, and this is indeed what subjects prefer.

In tasks 4 and 5 the situation is very different, now the payoffs of Person 3 grow in opposite direction of the payoffs of Person 1 creating a conflict between efficiency and maximin preferences. This is reflected in the two norm functions: while the linear utility model prescribes the choice of the most efficient allocation, the log utility model instead prescribes the maximin choice. We would like to emphasize at this point that we do not intend to suggest that only one of the two utility models is "correct." Rather, we see them as two ways of thinking about appropriateness of a given situation. One way of thinking considers only payoff differences and, thus, concludes that the efficient allocation is the most appropriate. The other takes into consideration the payoff differences *relative to wealth*, as is captured by diminishing marginal utility in the log utility model. This leads to higher weights on the dissatisfaction of "poor" Person 3 and, as a result, to the maximin choice being labeled most appropriate. Both ways of thinking may be reflected in subjects' behavior: roughly half choose the efficient allocation, with the other (roughly) half choosing the maximin. Interestingly, Baader and Vostroknutov (2017) found that students who chose to study economics and related subjects are more likely to maximize efficiency, while students in fields like European Studies and Arts and Culture prefer maximin. □

This case illustrates how the set of payoffs impacts normative valences, but also how the normative evaluation depends on how one values the payoffs. The potential for disagreement about norms stemming from different assessments of dissatisfaction (i.e., from different preferences over outcomes) merits further research. Next, we turn to a recent study by Galeotti et al. (2018) which analyzes the efficiency-equality trade-off in bargaining situations.

Case 2. Galeotti et al. (2018). In the experiment subjects chat in pairs about choosing between two or three allocations for themselves. Since there is no single decision maker and both subjects must agree on a choice, there is reason to think that the influence of norms will be particularly strong here, which should make the agreed-upon option more in line with social appropriateness. That said, the subjects have two minutes to negotiate, and if they do not reach an agreement, both get nothing. This feature can still lead to more aggressive subjects' achieving the consequence with higher material payoff for themselves.

Some of the tasks consist of allocations (x, x); $(120, 40)$; $(40, 120)$ where $x \in \{30, 40, 50, 60, 70, 80\}$. Thus, one allocation gives equal number of points to the negotiators and the other two give unequal numbers, but with the property that the unequal allocations are (weakly) more payoff efficient than the equal one. Figure 3 shows the difference in percentages of equal and unequal choices (solid line). We see that when x is small subjects choose unequal, but efficient, allocations. When x is large enough the modal choice switches to the equal allocation, at some cost to efficiency.

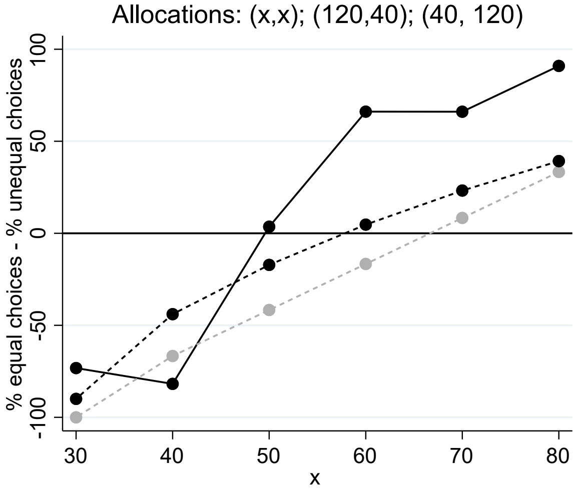


Figure 3: Solid line represents the data from Galeotti et al. (2018). Dashed lines show the predictions implied by norm functions computed under both the linear and the log utility models (grey and black lines respectively).

To compare the predictions of our model with these data, we again compute two variants of aggregate dissatisfaction for each allocation: with linear and log utilities over payoffs. The dashed black line on Figure 3 shows the differences in dissatisfactions between the equal and unequal allocations under the log utility model (the scale on the y-axis is arbitrary but the zero is set at the same level as zero for the differences in percentages). For the positive differences the model predicts that the equal allocation is the most appropriate and for negative differences that the unequal allocation is the most appropriate. We see that this prediction is in line with the choice of the majority of subjects (except for $x = 50$). The grey dashed line shows the differences in dissatisfactions computed under the linear utility model. In this case, linear utility does worse than log utility in accounting for behavior. □

This case shows that our model can capture the efficiency-equality trade-off studied in Galeotti et al. (2018). More importantly, the relative magnitudes of the dissatisfactions calculated for equal and unequal allocations can predict whether subjects prefer equality or efficiency. In particular, if the dissatisfactions are very similar, as, for example, in case $x = 50$ on Figure 3, then some pairs of subjects will converge to choosing equality and some efficiency. It is not surprising that when the normative valences of the two outcomes are very close to each other, choices are more variable. This case also demonstrates that our model, unlike standard social utility specifications, can be easily applied to unstructured bargaining environments where there is no one person who decides but where all players must come to a mutual agreement about the choice.

3.2 Expanding and Contracting the Set of Consequences

In this section we consider a set of studies in which the experimental manipulation adds or removes some consequences. In our model, this changes the normative valences of the consequences that are present in both cases, and thus can change the behavior.

We start with the give and take DGs analyzed by Bardsley (2008), List (2007), and Cappelen et al. (2013), among others. In these studies, it has been shown that subjects' generosity in the DG decreases when an additional action is added to an otherwise standard dictator game, allowing the dictator to take some money from the recipient. Since we do not observe the distributions of ϕ_i in the give-take experiments we assume that it is the same for all treatments of a given study, and we check whether the changes in norms between treatments are qualitatively reflected in the behavior.

Case 3. List (2007). In the Baseline treatment of List (2007) all dictators have \$5 and choose how much of it to give to the recipient. The Take1 treatment is the same except there is an additional possibility to take up to \$1 from the recipient (all subjects have endowments, such that recipients still receive a positive payoff, even when the dictator takes). The same goes for the Take5 treatment (can take up to \$5).

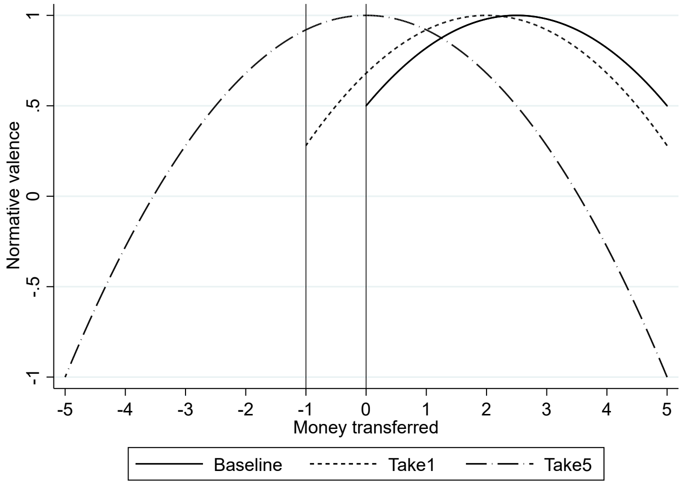


Figure 4: Relative norms in the three treatments of List (2007).

The graph on Figure 4 shows the *relative* norm functions in the Baseline, Take1, and Take5 treatments.[21] In all three cases, the aggregate dissatisfactions are calculated in the same way as for the standard DG (see Example 2). As we have proved in Proposition 2, the most socially appropriate consequence in a constant-sum, two-player game is the one that lies in the middle of the interval of monetary amounts that can be taken or given (when all consequences are equidistant). Thus, our model predicts that the most socially appropriate consequence involves less and less generosity as we move from Baseline to Take1 and to Take5. This is what List (2007)

[21] We renormalize the aggregate dissatisfactions in order for them to be comparable. Appendix C defines relative norm functions and explains how renormalization is done.

reports (see Figure 17 in Appendix A.1): in the Baseline treatment there is a spike at \$2.5 and in the Take5 treatment a spike at \$0 as predicted by our model. In the Take1 treatment, offers are less generous than in the Baseline treatment, however, there is no clear spike at \$2. Notice also that here, as compared to the previous section, the selfish motive of the dictator is given free reign, and thus many subjects choose to maximize their own payoff. As Kimbrough and Vostroknutov (2016, 2018) explain, this can be attributed to heterogeneity in the rule-following propensity: some subjects suffer high disutility from breaking norms (large coefficient ϕ_i in the utility, see Section 2.3), and some do not (low ϕ_i). □

Additional experiments with restricted giving options are needed to test the implications of our model for norms in Dictator games more thoroughly. Cox et al. (2018) report DG experiments with restricted giving options, along these lines. In most of their treatments the average offers are very close to our predictions, namely, the middle of the interval of possible consequences. Unfortunately we cannot say more since no other statistics are reported.

In the rest of this section we analyze extensive form two-moves games in which some consequences are removed. It is instructive to compare our theory with models of *reciprocal kindness* that attempt to explain behavior in dynamic games (Rabin, 1993; Charness and Rabin, 2002; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006); in doing so, we first relate our model to the conceptual discussion in Isoni and Sugden (2018) that highlights some philosophical difficulties that arise in reciprocity models of this kind.

Case 4. Isoni and Sugden (2018). In this paper the authors (IS) do not report any experiment, but rather analyze a simple two-move game shown in Figure 5. IS consider an ideal Trust World in which Player 1 chooses *send* and Player 2 chooses *return*, both with probability 1, while at the same time Player 2 chooses *equal* with probability less than 1 in the restriction of this game without the move of Player 1 (the game on the right). According to IS the idea of trust and trustworthiness is that in the game on the left Player 2 chooses *return* with higher probability than she is choosing *equal* in the game on the right exactly *because* Player 2 enters a trust relationship with Player 1 when he chooses *send*.

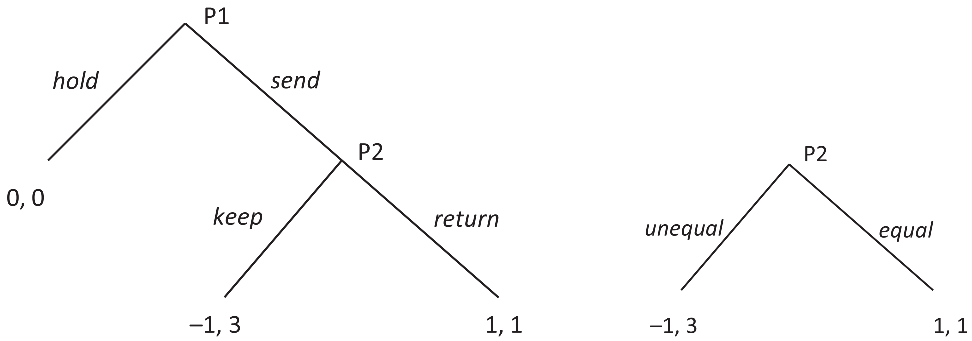


Figure 5: **Left:** the Trust game considered in Isoni and Sugden (2018). **Right:** the Dictator game faced by the second player in the absence of the move of the first player.

IS note that the models of reciprocal kindness by Rabin (1993), Charness and Rabin (2002), and Falk and Fischbacher (2006) do not support the above strategies as an equilibrium, while the model by Dufwenberg and Kirchsteiger (2004) does support it but fails to do so in other similar games. They call this the Paradox of Trust. The reason for the inability of these models to account for trust in this basic game lies in the way reciprocity is modeled: players are assumed to respond with kindness to kindness of other players, however, the action *send* does not classify as either kind or unkind since Player 1 chooses it *expecting* that Player 2 chooses *return*. IS conclude that trust behavior in this game cannot be based on reciprocal kindness, which presumes some reaction of Player 2 to some intentions of Player 1, and that this type of trust should instead be thought of as a "joint action" of the players who are involved in "reciprocal cooperation" (Isoni and Sugden, 2018).

Our theory of norms works exactly as IS suggest. The consequence $(1, 1)$ in the game on the left is the most appropriate, so the players who care enough about following norms do choose *send* and *return* in a "joint enterprise," which is to behave in the socially appropriate way. In the Dictator game on the right the two actions of Player 2 have the same normative valences, so, according to the norm-dependent utility, she chooses the selfish option *unequal*, exactly as IS hypothesize. Thus, our theory resolves the Paradox of Trust that is inherent in the models of reciprocal kindness and shows that trust can be based on social norms, which *create* reciprocal behavior (see Kimbrough and Vostroknutov (2021) for more discussion). □

Next, we analyze experimental findings of McCabe et al. (2003) that are consistent with the ideas of Isoni and Sugden (2018). This is the simplest extensive form game that allows us to test our model, since in the Trust game used by McCabe et al. (2003) the most appropriate action lies in the subgame, and thus the second mover should not punish the first for violating the norm. This feature makes it possible to look exclusively at the behavioral changes brought about by the removal of one consequence.

Case 5. McCabe et al. (2003). The authors (MRS) consider a simple trust game and its subgame played in separate treatments between-subjects (on the left of Figure 6).

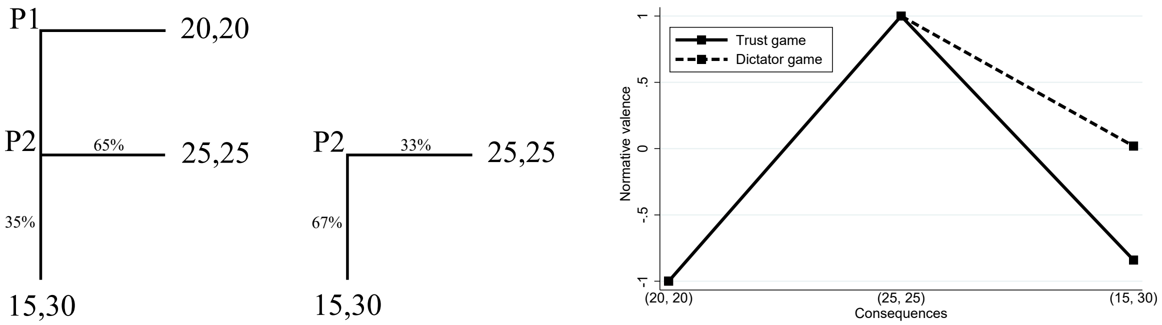


Figure 6: **Left:** the Trust and Dictator games analyzed in McCabe et al. (2003). **Right:** the normative valences of the consequences in the two games (log utility, relative norms). For the linear utility norms see Figure 18 in Appendix A.2.

MRS notice that the behavior of P2s depends on whether P1 moved first or not. Specifically, after the move of P1, 65% of P2s choose the cooperative consequence $(25, 25)$, while without this move 67% of P2s choose the selfish option $(15, 30)$. MRS explain this treatment difference with the idea that P2s want to reciprocate the trustful move of P1 and thus choose the cooperative option $(25, 25)$; while, without this move of P1 there is nothing to reciprocate, so more P2s choose selfishly.

According to our theory, this change in behavior follows from the different normative valences of consequences in the full Trust game and the associated Dictator game. The graph on the right of Figure 6 shows the norm functions calculated with log utility and renormalized relative to each other (see Appendix C). The results are qualitatively the same with linear utility (see Figure 18 in Appendix A.2). The normative valence of the consequence $(15, 30)$ is very low in the Trust game, but is around 0 in the Dictator game. Thus, the material payoffs for P2 are the same in the two games, but the difference in normative valences between the cooperative and selfish actions decreases in the Dictator game. Therefore, according to the norm-dependent utility, subjects with intermediate propensity to follow norms should switch from choosing cooperative action in the Trust game to selfish action in the Dictator game, exactly what the data suggest. □

In the case above, second movers choose to cooperate in the Trust game (consequence $(25, 25)$) because the existence of a forgone option (consequence $(20, 20)$) makes the appropriateness of the selfish option $(15, 30)$ much lower, so norm-following individuals avoid it. This shows how behavior in extensive form games can change due to expanding or contracting the set of possible outcomes.

3.3 Punishment

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 consequence, 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 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 6. Charness and Rabin (2002). The authors (CR) study the games shown in Figure 7 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.

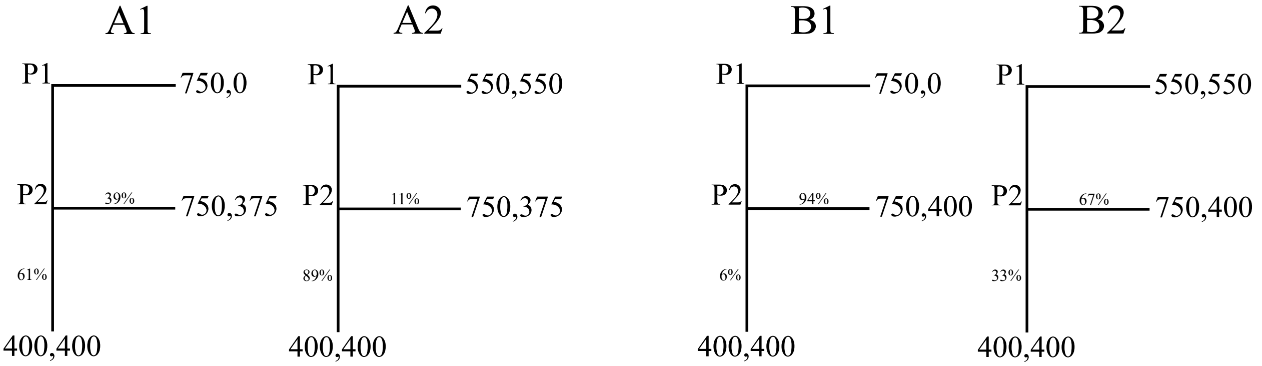


Figure 7: The games analyzed in Charness and Rabin (2002). In this study A1 is coded Berk21; A2 is combined Barc1 and Berk13; B1 is Barc7; and B2 is Barc5.

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 consequence, but in A1/B1 no norm is violated when P1 continues the game and so no punishment is expected.[22]

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 consequences $(750, 375)$ and $(750, 400)$ in the A and B games respectively have higher appropriateness than the consequence $(400, 400)$, which is consistent with the observation that a non-negligible number of subjects choose these options. □

Case 6 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

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

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 7. 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.

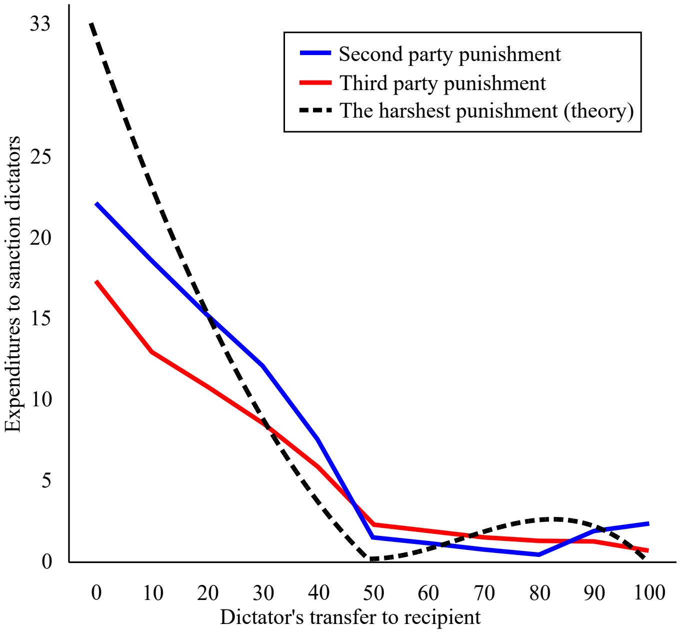


Figure 8: Third and second party punishment in the DG reported in FF. The dashed line shows the model predictions of the harshest possible punishment meted by the extreme norm-followers with very little costs of punishment.

Figure 8 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 an Eye for an Eye (EE) 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 EE, 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.[23]

Finally, in Case 12 in Appendix A 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). □

4 Norms in Context

In this section, we show how our model can be generalized to account for a number of fundamental features of human social life, including respect for ownership claims and role entitlements, expression of in- and out-group distinctions, deference to social status, and kin favoritism. These are human universals, present to varying degrees in all societies, and thus we take them as axiomatic and merely show how they might be accounted for within the structure of our model of injunctive norms. We do so by altering the way we handle aggregation in the model. Notice that in the definitions in Section 2 all players and their payoffs are treated equally in aggregation. This represents a "baseline" model of norms in interactions between symmetric, co-equal agents. As noted in the introduction, the baseline model neatly captures the situation in typical lab experiments, in which all subjects belong to the same group of people (students), have indistinguishable social status (because of anonymity), lack role entitlements and ownership claims, and make decisions regarding the allocation of windfall resources provided by the experimenter. We show what happens in the model when we break this symmetry in several different ways. The derivation of norm-dependent utility in general games with observable actions for the norm in context is explained in Appendix D.3.

4.1 Ownership Claims

The endowment effect (Thaler, 1980) shows how the perception and valuation of an object can change when some form of ownership over it is established. Ownership claims also influence the sharing rate in social dilemmas (e.g., Gächter and Riedl, 2005; List, 2007; Oxoby and Spraggon, 2008). When subjects believe that they own the resources that they are asked to share, they tend to be more selfish than when the resource is unowned (or owned by someone else). We conjecture that ownership claims are reflected in social norms. This sounds intuitively plausible, as situations in which people refuse to give away things that they own do not seem inappropriate. In what follows, we make a distinction between *ownership claims*, which refer to some resources

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

being owned by an individual, and *role entitlements* (discussed in Section 4.2 below) which refer to a person's entitlement to some position and an associated set of actions (e.g., with respect to the allocation of some unowned resources).

We model ownership claims by assuming that possible (re)allocations of owned resources trigger dissatisfaction differently from the *windfall payoffs* that are typical in experiments. We assume that the utility of owned money is the same as the utility of unowned money. What changes is the intensity of feelings of dissatisfaction related to *losing* the money of the former type. To capture this in the model, we assume that each amount of money that a player might receive is divided into several pieces, which differ in their degree of ownership by the players. Mathematically, we redefine the utility function from the previous section to be $u : C \to \mathbb{R}^{NP}$, where P is a finite set of separate ownership classes (in Section 2, P consisted of one element). Thus, player i derives utility $u_{ip}(x)$ from the ownership class p in consequence $x \in C$. Finally, let $\pi_{ip} \in [0, 1]$ denote the *ownership weight* that is assigned to player i in class p; this is intended to capture the strength of player i's ownership claim over the resources in p. For each class, the weights determine the distribution of ownership of the resources in it. In general, we require that if p is an ownership class, then $\sum_{i \in N} \pi_{ip} = 1$.

For example, suppose that player i "completely" owns the resources in class p. Then $\pi_{ip} = 1$ and $\pi_{jp} = 0$ for all other players $j \neq i$. Player i might also have a weaker ownership claim. In this case we can have $\pi_{ip} = 0.8$, $\pi_{kp} = 0.2$, and $\pi_{jp} = 0$ for all $j \neq i, k$. Here players i and k own the resources together, but have different "shares," like partners in a firm. We retain windfall payoffs as the special case p' in which $\forall_{i,j \in N}\, \pi_{ip'} = \pi_{jp'}$.

To introduce ownership claims to the definition of a norm function we update the dissatisfaction formula above to

$$d_i(x, c) := \max\{\sum_{p \in P} \pi_{ip}(u_{ip}(c) - u_{ip}(x)), 0\}. \tag{4}$$

To calculate the dissatisfaction for x because of c we first sum up the utility differences weighted by the ownership weights in all classes. Notice that we do not require that only positive differences count *in each class*, but that negative differences can counterbalance positive ones. This seems reasonable since, in the end, we are talking about a single set of resources, even though different pieces of it have different associated ownership claims. This formulation allows for interesting cases in which a player has some fixed amount of money in each of two consequences, but the ownership claim over this money changes. In this case she will be dissatisfied with the consequence in which her ownership is decreased.

This is the only modification we need in order to incorporate ownership claims. The rest of the definitions stay unchanged. We proceed with some examples that demonstrate how ownership of resources can change norm functions in allocation decisions and social dilemmas.

Example 3. DG with Ownership Claim. Suppose a dictator p is asked to share his own hard-earned money with a stranger r. To analyze this situation, we extend the analysis in Example 2 by introducing one ownership class with $\pi_p = 1$ and $\pi_r = 0$ (we drop the subscript for the class since there is only one). Notice that here we do not have in mind an experiment à la Hoffman et al. (1994), where the *right to be* a dictator is earned through a contest—this possibility is considered in Section 4.2 below—but rather a situation in which subjects bring their own money to the lab and are asked to use it in a dictator game. The aggregate dissatisfaction in this case is

$$D(c) = D_p(c) + D_r(c) = \frac{c^2}{2} + 0.$$

Thus, when the dictator owns the entire pie, the most socially appropriate consequence is to give the receiver nothing. In general, for arbitrary π_p and π_r with $\pi_p + \pi_r = 1$, the most socially appropriate consequence is $c^* = 1 - \pi_p = \pi_r$; the dictator ought to give the receiver whatever share of the pie she owns. □

We illustrate how this works via a set of experiments that manipulate ownership claims. Cherry et al. (2002), Oxoby and Spraggon (2008), and Korenok et al. (2017) study Dictator games where dictators or recipients earn money by answering questions from the GMAT, thus inducing an ownership claim. The results in these studies are similar, so we focus on the most comprehensive one. It should be mentioned that we do not have a theory about how exactly the ownership claims are established. We simply assume that having put sufficient effort into earning money creates a feeling of ownership.[24]

Case 8. Oxoby and Spraggon (2008). In the treatments DE and RE of Oxoby and Spraggon (2008) only dictators or only recipients earn money that are later divided between the two players by the dictator. In the Baseline treatment the money is assigned randomly by the experimenters. Subjects can earn \$10, \$20, or \$40 depending on how many GMAT questions they answer correctly (0-8, 9-14, or 15-20 questions). Thus, earning \$20 or \$40 signals effort and creates an ownership claim.

We model this situation by assigning an ownership claim weight 1 to the payoff of the dictator or the recipient. In the Baseline treatment we assume windfall payoffs with weights $\frac{1}{2}$. The left

[24] John Locke in his theory of natural law (Locke, 1690) states that individuals deserve property entitlements in resources that they acquired through their own expenditure or labor. This coheres with many peoples' intuition and may explain the attractiveness (to some) of the labor theory of value. Nevertheless, some evidence suggests that not just any effort is sufficient to create a strong sense of ownership: some studies that use tedious or menial tasks that anyone can perform fail to detect any influence of earning one's money in this way on behavior (e.g., Cappelen et al., 2013). Thus, we propose that the reader always considers the possibility that in some experiments the tasks designed to create feelings of ownership might have failed to do that.

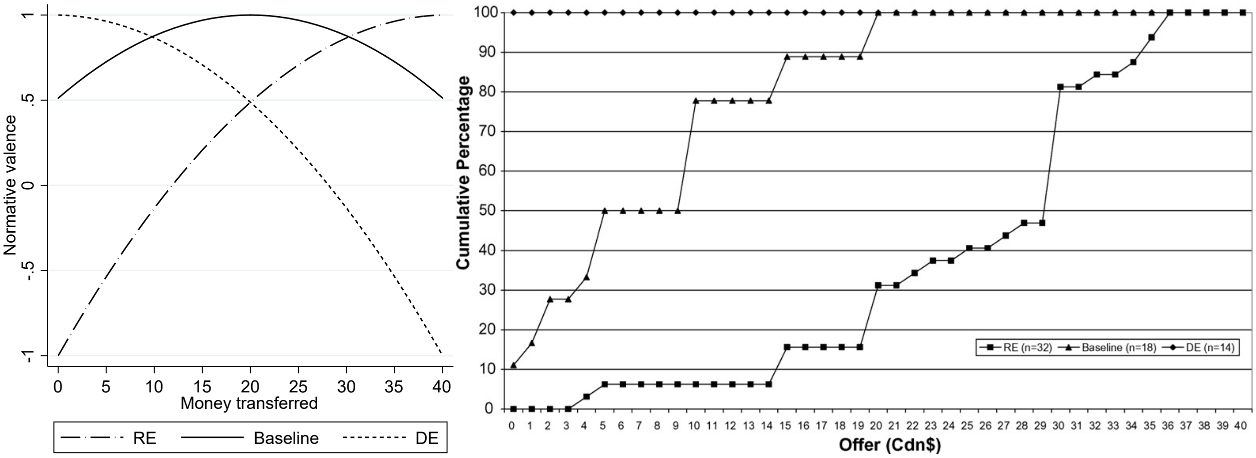


Figure 9: **Left:** relative norm functions in the three treatments of Oxoby and Spraggon (2008): Recipient Earnings (RE), Baseline, and Dictator Earnings (DE). **Right:** cumulative distributions of offers in these treatments.

graph on Figure 9 shows the resulting norm functions. When the dictator earned the money it is most appropriate to give nothing to the recipient; when the recipient earned the money it is most appropriate to give her everything; and in the Baseline treatment the money ought to be divided equally. The right graph on Figure 9 shows the cumulative distributions of offers in the three treatments for subjects who earned $40. Notice that in the DE treatment all dictators keep all the money as predicted by the norm. In the RE treatment 63% of dictators offer more than half of the money, which is remarkable in comparison with the Baseline treatment where no one offers more than half. This clearly demonstrates the effect of the ownership claim of the recipient. Similar results are obtained for the case when subjects earn $20 (Figure 19 in Appendix A.3). □

Example 3 above represents the simplest case because there is only one ownership class. However, there are many important situations in which different people own various inputs to various degrees (i.e., there are multiple ownership classes); in these cases, the intuition remains that the appropriateness of giving someone a resource is increasing in their ownership claim. One implication is that, under the model, one can use treatment variation to estimate (the strength of) perceived ownership. Our next analysis is presented in Case 11 in Appendix A. We consider again the experiment by List (2007) that we focused on in Case 3 above. In one of the treatments (Earnings) both dictators and recipients put effort into earning their endowments, which makes it different from Oxoby and Spraggon (2008) where only one subject earns it. We model this situation with two ownership classes. The majority of subjects do not give any money to the recipient and do not take any money from her either. This is consistent with our model if both parties have a complete ownership claim to their endowments.

4.2 Role Entitlement

Role entitlements differ from ownership claims in that an individual with a role entitlement does not necessarily own the money that her role entitles her to control. For example, a bureaucrat may be empowered to distribute some public funds without having ownership over them. However, he has an authority to decide how the funds should be divided, and thus others without the role entitlement have limited scope to disapprove of his decisions. In their famous experiments, Hoffman and Spitzer (1985) and Hoffman et al. (1994) explicitly state that the right to a role in a game implies "the guarantee . . . against reprisal" (Hoffman et al., 1994).[25]

The example of the bureaucrat highlights how we incorporate role entitlements; we assume that such entitlements do not alter the dissatisfaction of the players with the various outcomes, as was the case with ownership claims, but rather that role entitlements change the resentment associated with norm violations. We thus propose that if a player is entitled to some role, then he is resented less for taking socially inappropriate actions. Thus, role entitlements change the intensity of punishment, which is expressed through a weight $\sigma \in [0, 1]$ in Section 2.2. We extend the model by assuming that role entitlements are associated with different weights for different players. Player i with a role entitlement has parameter $\sigma_i \in [0, 1]$, with $\sigma_i = 1$ meaning that i has the strongest entitlement, and he should not be punished at all; $\sigma_i = 0$ means no entitlement and the punishment should be full. Thus, when there exists an exogenous punishment mechanism, this means that the punishment norm function is $\sigma_i + (1 - \sigma_i)\mu_i(c)$, a convex combination of full appropriateness and the punishment norm function. When there is no separate punishment mechanism, the norm function with punishment becomes $\eta'(c) = \sigma_i \eta_C(c) + (1 - \sigma_i)\mu_i(c)$ as in Section 2.2.

Case 9. Hoffman et al. (1994). In the classic experiment by Hoffman et al. (1994) subjects play the $10 UG with either random assignment of roles or role assignment determined by a contest (general knowledge quiz, winner becomes the proposer). The authors observe that the mode of the distribution of offers shifts from $5 in the random assignment treatment to $4 in the role entitlement treatment (the difference is significant). The authors do not report the average offers, but in a recent replication by Fleiß (2015) with the pie size of $20, the average offer significantly decreases from 7.64 in the random assignment treatment to 6.48 in the role entitlement treatment.

In our model, this shift in offers can be interpreted as an increase in the σ_p coefficient of the proposer with role entitlement, which makes the weight on the punishment function smaller. Figure 10 shows the norm functions with random role assignment (thin grey and red lines, $\sigma_p =$

[25] It should be noted, however, that the terminology of Hoffman and Spitzer (1985) and Hoffman et al. (1994) is somewhat different from ours. The authors of these studies, when talking about subjects' earning the right to be a proposer in the Ultimatum game, refer to "property rights" and the Lockean theory of desert. According to them, becoming a proposer through a contest entitles a subject to ownership of the pie. In our model we draw a distinction between earning the money by working on some task (ownership claim) and earning the right to be a proposer (role entitlement).

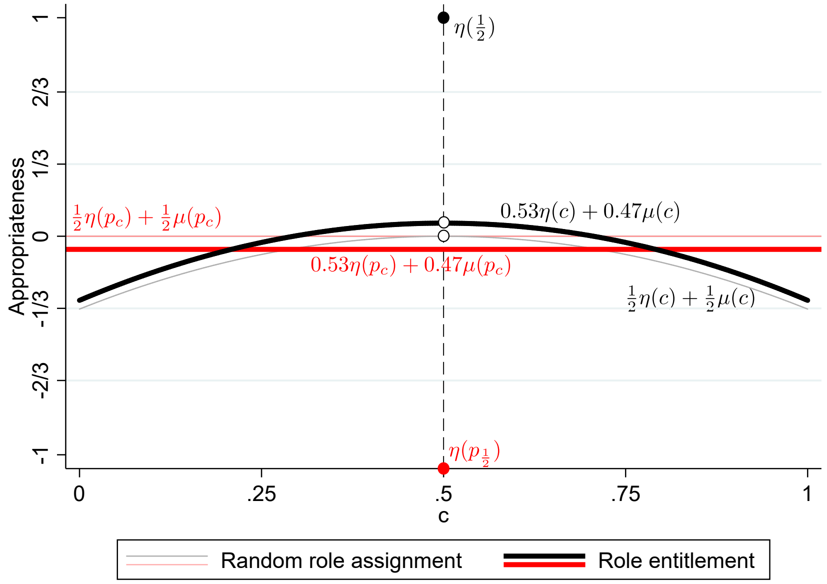


Figure 10: Norm functions in the UG with random role assignment and role entitlement.

$\frac{1}{2}$) and role entitlement (thick black and red lines, $\sigma_p = 0.53$).[26] The appropriateness of accepting relatively unequal offers goes up, and the appropriateness of rejecting goes down. In the vicinity of half-half division, the role entitlement of the proposer makes the consequences, in which the offer is accepted, more appropriate than those where it is rejected, as compared to the random entitlement case where acceptance is always less appropriate than rejection. Thus, a strategic proposer should offer less when entitled to the role due to the lower likelihood of rejection.

Under the model, the same offers should be rejected less often in the role entitlement treatment. Hoffman et al. (1994) observe very few rejections in both treatments. This does not contradict the model, but does not support it either. However, Fleiß (2015) observes a significant drop in acceptance thresholds of receivers (using the strategy method) from 6.51 to 4.73 ($20 pie), which is in line with our model.[27] □

4.3 Discrimination: Social Status, Kinship, and In- and Out-group

Discrimination, in the form of differential treatment of in- and out-group members, deference to high social status individuals, and favoritism toward kin, is ubiquitous in human societies, and these tendencies are well documented (e.g., Brown, 2000; Buss, 2005). As in the previous sections, we take the existence of such distinctions as given and model them in our framework by means of weights similar to those used to model ownership claims. The difference is that

[26] To clarify notation, check Example 5 in Appendix A that describes standard UG.

[27] It should be mentioned that at least one study failed to replicate the findings above (Demiral and Mollerstrom, 2018). We think that one reason for this could be the task that the authors used to induce role entitlement. Instead of a general knowledge quiz, as in Hoffman et al. (1994) and Fleiß (2015), they used a number summation task. There may be variation in the perceived legitimacy of a role entitlement; perhaps a task that "anyone can do" does not induce strong perceptions of entitlement.

in case of discrimination the dissatisfaction of individual players gets amplified or diminished not because of its source (ownership), but because of *who* the players are. For example, the dissatisfaction of a player with high social status has more weight than the same dissatisfaction of a low status individual. For simplicity, such weights are assumed to be part of the "culture" and to be commonly recognized by all players. This would imply, for example, that transferring wealth to someone with relatively higher social status is considered appropriate. With the out-group the situation is similar: the dissatisfaction of everyone who belongs to the out-group is downgraded with a common weight, so acting selfishly with the out-group is considered more appropriate than doing so with the in-group. With kin we assume that the dissatisfaction of a related individual is weighted proportionally to the degree of relatedness (nuclear family, extended family). This implies that selfish behavior is increasingly appropriate towards more unrelated individuals.

Suppose the set of players N is partitioned into two groups $N = \{N_1, N_2\}$. Each player i has a weight $\tau_i \in [0, 1]$ that defines her relative status among the players in her group, with higher weight implying higher status. The dissatisfactions of the out-group players are discounted with a weight $\rho_{k\ell} \in [0, 1]$ where k is the index of the in-group and ℓ is the index of the out-group ($k, \ell \in \{1, 2\}$). In principle, $\rho_{k\ell}$ can be allowed to be negative; this would model outright hostility towards the out-group. We then apply the weights when computing aggregate dissatisfaction across individuals to generate the norm function. However, we first define two norm functions, one from the perspective of members of each group.[28] The aggregate dissatisfaction from the perspective of group N_1 is

$$D(x|N_1) := \sum_{i \in N_1} \tau_i D_i(x) + \rho_{12} \sum_{i \in N_2} \tau_i D_i(x), \tag{5}$$

and symmetrically from the perspective of group N_2:

$$D(x|N_2) := \sum_{i \in N_2} \tau_i D_i(x) + \rho_{21} \sum_{i \in N_1} \tau_i D_i(x).$$

These definitions can be easily generalized to any number of groups with a complex system of relationships defined by group-specific coefficients $\rho_{k\ell}$.[29]

[28] The idea that norms are indexed to groups has been employed elsewhere; see Pickup et al. (2019) and Chang et al. (2019), who suggest that those who share a particular group identity are aware of and adhere to norms associated with that identity.

[29] With status weights τ_i, it is possible to hypothesize what they might depend on. One candidate is the norm-following parameter ϕ_i. Indeed, in most societies, people with, say, permanent criminal record (low ϕ_i) do not have access to the same jobs and opportunities as people without, which signifies that it is appropriate from the societal perspective to treat such individuals as undeserving (low status). Symmetrically, people who, for example, choose to serve in the army are respected and provided with privileges, because by choosing to put their lives in danger for the society they unambiguously signal their high ϕ_i. Thus, status weight τ_i can in principle be increasing in *beliefs* about someone's ϕ_i, which paves the way to normative theories of social image and reputation concerns.

Finally, for each player $j \in N$ we define kin relationships by means of weights $\kappa_{ji} \in [0,1]$, where $i \in N$ indexes all other players. We assume that $\kappa_{jj} = 1$ for all $j \in N$ (own dissatisfaction is counted as the most important) and that κ_{ji} are connected to the degree of relatedness. For example, κ_{ji} can be taken to be proportional to the weights defined by Hamilton's law (Hamilton, 1964), though genetically unrelated individuals like spouses and their kin can also have relatively high weights. With kin relationships the norm functions become different for each individual j. We define the general aggregate dissatisfaction of player j belonging to group N_1 as

$$D(x|N_1, j) := \sum_{i \in N_1} (\tau_i + \kappa_{ji}) D_i(x) + \rho_{12} \sum_{i \in N_2} (\tau_i + \kappa_{ji}) D_i(x). \tag{6}$$

Definition (6), which can also be easily generalized to any number of groups, incorporates all "community relevant" information into the normative valence of each consequence. It should be noted that we propose a very simple functional form $\tau_i + \kappa_{ji}$ for the relative importance of kinship and social status. Undoubtedly, other ways of combining weights are possible, but which way is best is an empirical question.

Evidence on Status and Kinship. Under the model, the dissatisfactions of individuals with higher social status should receive higher weights in the aggregate dissatisfaction function D. This implies that in the same situation higher status individuals should enjoy higher payoffs than those with lower status. The evidence of status deference is aplenty in anthropology, human evolutionary biology, and evolutionary psychology (see e.g., Cummins, 2005). In the economics literature, Ball et al. (2001) find that high status subjects earn a disproportionate share of the gains from exchange in experimental double auction markets that employ a box design, in which demand equals supply for a continuum of prices. Prices favor the high status side of the market, whether status is thought to be "earned" through performance on a quiz or randomly assigned. Similarly, if individuals weight the dissatisfaction of their kin according to their degree of relatedness, the theory predicts that the appropriateness of kindness toward kin is increasing in it. Madsen et al. (2007) show that when subjects are told that their kin will be paid based on the amount of time that the subjects spend performing a painful exercise known as a "wall squat," their willingness to tolerate pain is increasing in the closeness of the kin relationship (i.e., people will suffer longer for a brother than a cousin). Both of these results can be interpreted as high status individuals' and kin's dissatisfaction receiving higher weights.

Evidence on In- and Out-Group Discrimination. We analyze a study that employs the minimal group paradigm from social psychology (Tajfel and Turner, 1986) to induce in- and out-group identities. This experiment is particularly useful for our purposes because it employs a within-subject design which allows us to estimate the weights on in- vs. out-groups from an allocation task and ask how well those weights (and the implied norms) predict play in a subsequent series of games.

Case 10. Chen and Li (2009). The authors (CL) use the classic Klee-Kandinsky method to assign individuals to groups and strengthen their identification with those groups with some additional tasks that include, for example, a chat with an in-group member about the characteristics of the paintings. Then subjects choose how to allocate tokens between two other subjects (other-other task) who either both belong to the in-group, both belong to the out-group, or one of each (decision makers are not incentivized). Figure 11 shows the results.

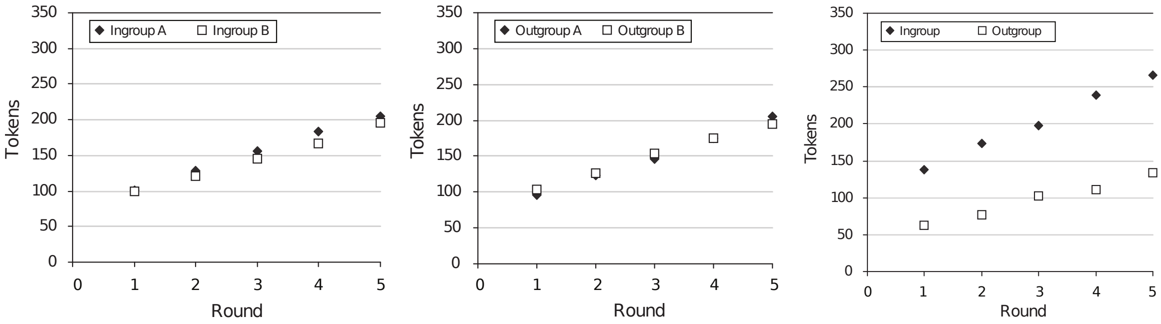


Figure 11: Choices in the other-other allocation tasks of Chen and Li (2009) with different compositions of others.

When subjects allocate tokens across two people from the same group (only in-group or only out-group) they divide the tokens equally, as predicted by our model of the Dictator game with equal weights on the dissatisfactions of the players. However, when one recipient is an in-group member and the other is an out-group member, subjects favor the in-group at a ratio of 2:1 for each of the five rounds with different endowments. This observation can be rationalized with a weight on the out-group equal to $\rho = \frac{1}{2}$, which implies that subjects treat the dissatisfaction of in-group members as twice as important as those of the out-group. We can use this weight to test the comparative statics of the theory using data from the second part of the experiment.

After this task subjects played in a sequence of 23 one- or two-moves games taken from Charness and Rabin (2002). There are two conditions: in the first, both players are from the *same* group (in-group) and in the second, the players are from *different* groups (out-group). Games, choices, and norm functions are reported in Table 2 in Appendix A.4. The norm functions are computed with the weight $\rho = \frac{1}{2}$ that we estimated from the other-other tasks. To assess how well our model can account for observed changes in choice proportions between the in- and out-group games, we focus on second movers, since the first mover's behavior depends on the beliefs about what the second mover will do and on many unknown parameters. Second movers always choose between Left and Right. We compute variables ΔChoice and ΔNorms (last two columns of Table 2). The former is the difference in the proportion of choices of Left of players B (second movers) between the out-group games and the in-group games. The latter is the difference of differences between the out-group and in-group games of the norms associated with choices Left and Right of player B.[30] This quantity measures the change in norm-dependent utility of the second mover and should be proportional to the change in choices of Left if our model is correct.

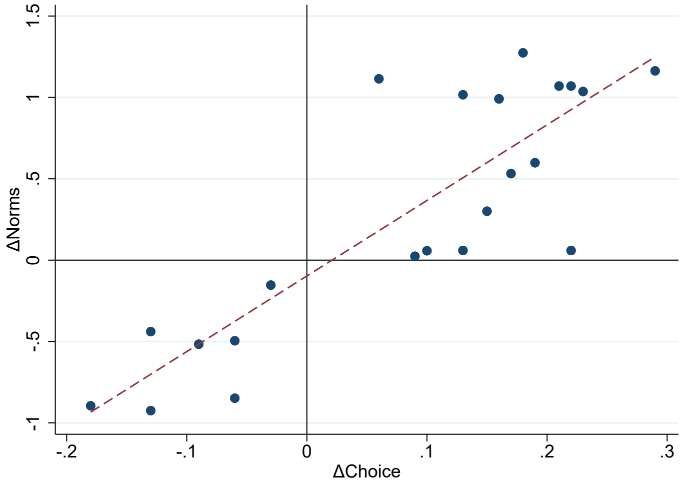


Figure 12: Change in choices and change in normative valences predicted by our model for the 23 games studied in Chen and Li (2009).

Our first observation is that in *all* games the change in choice has the same sign as the change in normative valences. Thus, our theory can account for the direction of change in all games. Moreover, if we assume a random utility specification as CL do, then the change in proportion of choices should be proportional to the change in normative valences, since they enter the norm-dependent utility. Figure 12 shows a scatter plot of the two variables. The dashed line is the OLS regression with robust errors ($\beta = 0.16$, $p < 0.001$). Spearman's rank correlation is 0.79 ($p < 0.0001$). This provides strong quantitative support to our theory. □

5 Conclusion

We propose the first (to our knowledge) theory of injunctive norms in games. The theory is intended to provide structure to models of norm-dependent utility, which have been shown to have substantial predictive/explanatory power in experimental games. This power has been made possible, in part, because theories of norm-following introduce additional free parameters making it easier to fit the data. This has raised concern that such models provide too many "modeler degrees of freedom," and so we have sought to address this concern by showing how a measure of normative appropriateness of each outcome can be defined only in terms of the set of possible outcomes in a game.

[30] Strictly speaking, there are 6 out of 23 games in which an action by the first mover will lead to different punishment functions for in- and out-group members. In what follows we ignore this when computing the change in norm functions; if we simply exclude those games instead, our results for the remaining 17 games are essentially identical.

The theory assumes that normative evaluations aggregate the emotional reaction of each interested party to the possible outcomes. We assume that normative evaluations are driven by comparative dissatisfaction when individuals evaluate counterfactual opportunities to earn higher payoffs. The normatively most appropriate outcome is the one that minimizes aggregate dissatisfaction of this kind. We take the theory to existing data to show how it can rationalize a variety of seemingly puzzling observations about social behavior, including the fact that measured social preferences are known to vary across contexts, the fact that adding/subtracting seemingly irrelevant outcomes to/from a game can change behavior, and the nature and intensity of costly punishment.

An important virtue of the model is the ease with which it can be applied. Computing the normative appropriateness of each outcome is straightforward, and then one need only use the appropriateness measure as an input to norm-dependent preferences, and the resulting game can be analyzed with standard tools.

While the evidence we present is largely consistent with the model, another key virtue of the theory is that it establishes a falsifiable framework for studying the influence of norms on behavior. Suitably designed experiments will thus be able to more thoroughly test the theory's implications and probe the boundaries of its applicability. We have little doubt that the present model is incomplete, but we view it as a valuable step in the right direction.

Finally, while the theory makes predictions about a number of important ways that the context in which a choice is made matters for behavior, the basic model is nevertheless unable to account for a variety of other "context effects" that have been documented in the literature, such as the effects of entitlements/ownership and the effects of individual and group status/identity on behavior. In the final two sections we show how the model can be extended in a straightforward way to provide an account of such observations. The key intuition is that ownership, entitlements, and discriminatory treatment can all be understood as factors that change the weights used in aggregation of dissatisfaction within or across individuals in constructing the norm function.

Of course, once we account for these factors, we once again introduce many parameters, specifically weights π that determine ownership claims, punishment weights σ for the role entitlements, and weights τ, ρ, κ for status, in/out-group, and kin relationships. If the model is taken at face value, this means that experimental designs can be used to infer these weights from data in order to help improve our understanding of normative variation. However, we also note that, even if the model is "correct enough" to be used in this fashion, there would remain substantial room for normative uncertainty and disagreement about what is appropriate, especially given that entitlements and social relationships may not be precisely determined. We think this source of normative uncertainty is a plausible source of conflict and that this is an important direction for future research.

References

Baader, M. and Vostroknutov, A. (2017). Interaction of reasoning ability and distributional preferences in a social dilemma. *Journal of Economic Behavior & Organization*, 142:79–91.

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

Ball, S., Eckel, C. C., Grossman, P. J., and Zame, W. (2001). Status in markets. *Quarterly Journal of Economics*, 116(1):161–188.

Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics*, 11:122–133.

Brown, D. E. (2000). Human universals and their implications. In Roughley, N., editor, *Being Humans: Anthropological Universality and Particularity in Transdisciplinary Perspectives*. New York: Walter de Gruyter.

Buss, D. M., editor (2005). *The Handbook of Evolutionary Psychology*. John Wiley & Sons, Inc.

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.

Cappelen, A. W., Nielsen, U. H., Sørensen, E. Ø., Tungodden, B., and Tyran, J.-R. (2013). Give and take in dictator games. *Economics Letters*, 118(2):280–283.

Chang, D., Chen, R., and Krupka, E. (2019). Rhetoric matters: a social norms explanation for the anomaly of framing. *Games and Economic Behavior*.

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

Chen, Y. and Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, 99(1):431–57.

Cherry, T. L., Frykblom, P., and Shogren, J. F. (2002). Hardnose the dictator. *American Economic Review*, 92(4):1218–1221.

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.

Cox, J. C., List, J. A., Price, M., Sadiraj, V., and Samek, A. (2018). Moral costs and rational choice: Theory and experimental evidence. mimeo, Georgia State University, University of Chicago, University of Alabama, University of Southern California.

Cummins, D. (2005). Dominance, status, and social hierarchies. In Buss, D. M., editor, *The Handbook of Evolutionary Psychology*, chapter 20, pages 676–697. John Wiley & Sons, Inc.

Demiral, E. E. and Mollerstrom, J. (2018). The entitlement effect in the ultimatum game–does it even exist? *Journal of Economic Behavior & Organization*.

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.

Engelmann, D. and Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution. *American Economic Review*, 94(4):857–869.

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.

Fleiß, J. (2015). Merit norms in the ultimatum game: an experimental study of the effect of merit on individual behavior and aggregate outcomes. *Central European Journal of Operations Research*, 23(2):389–406.

Gächter, S. and Riedl, A. (2005). Moral property rights in bargaining with infeasible claims. *Management Science*, 51(2):249–263.

Galeotti, F., Montero, M., and Poulsen, A. (2018). Efficiency versus equality in bargaining. *Journal of European Economic Association*, forthcoming.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. i. *Journal of Theoretical Biology*, 7(1):1–16.

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

Hoffman, E., McCabe, K., Shachat, K., and Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, 7:346–380.

Hoffman, E. and Spitzer, M. L. (1985). Entitlements, rights, and fairness: An experimental examination of subjects' concepts of distributive justice. *Journal of Legal Studies*, 14:259–298.

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

Isoni, A. and Sugden, R. (2018). Reciprocity and the Paradox of Trust in psychological game theory. *Journal of Economic Behavior & Organization*.

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. (2021). Axiomatic models of injunctive norms and moral rules. mimeo, Chapman University and Maastricht University.

Korenok, O., Millner, E., and Razzolini, L. (2017). Feelings of ownership in dictator games. *Journal of Economic Psychology*, 61(C):145–151.

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.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, 115(3):482–493.

Locke, J. (1690). *The second treatise of civil government*. Awnsham Churchill.

Loomes, G. and Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The economic journal*, 92(368):805–824.

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

Mackie, J. L. (1982). Morality and the retributive emotions. *Criminal Justice Ethics*, 1(1):3–10.

Madsen, E. A., Tunney, R. J., Fieldman, G., Plotkin, H. C., Dunbar, R. I., Richardson, J.-M., and McFarland, D. (2007). Kinship and altruism: A cross-cultural experimental study. *British Journal of Psychology*, 98(2):339–359.

McCabe, K. A., Rigdon, M. L., and Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization*, 52:267–275.

Merguei, N., Strobel, M., and Vostroknutov, A. (2020). Moral opportunism and excess in punishment decisions. mimeo, Maastricht University.

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.

Oxoby, R. J. and Spraggon, J. (2008). Mine and yours: Property rights in dictator games. *Journal of Economic Behavior & Organization*, 65(3-4):703–713.

Panizza, F., Vostroknutov, A., and Coricelli, G. (2020). How conformity can lead to extreme social behaviour. mimeo, University of Trento, Maastricht University, University of Southern California.

Panizza, F., Vostroknutov, A., and Coricelli, G. (2021). The role of meta-context in moral decisions. mimeo, Maastricht University, University of Trento, and University of Southern California.

Pickup, M., Kimbrough, E. O., and de Rooij, E. (2019). Expressive politics as (costly) norm following. SSRN Working Paper 2851135.

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

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, 83(5):1281–1302.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

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

Smith, V. L. and Wilson, B. J. (2017). *Sentiments*, conduct and trust in the laboratory. *Social Philosophy and Policy*, 34(1):25–55. Economic Science Institute Working Paper.

Stigler, G. J. and Becker, G. S. (1977). De gustibus non est disputandum. *The American Economic Review*, 67(2):76–90.

Sugden, R. (2018). *The Community of Advantage: A Behavioural Economist's Defence of the Market*. Oxford University Press.

Tajfel, H. and Turner, J. (1986). The social identity theory of intergroup behavior. In Worchel, S. and Austin, W., editors, *The psychology of intergroup relations*, pages 7–24. Chicago: Nelson-Hall.

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, 1(1):39–60.

Tremewan, J. and Vostroknutov, A. (2021). *A Research Agenda in Experimental Economics*, chapter An Informational Framework for Studying Social Norms. Edward Elgar Publishers.

Appendix (for online publication)

A Additional Examples, Cases, and Supporting Evidence

Example 4. Adding Replicates of the Same Consequence Matters. Augmenting the set of consequences with additional consequences that introduce new payoff possibilities can clearly impact the injunctive norm under the model. However, one striking implication of our theory is that even adding a copy of an existing consequence (such that the set of possible final payoff vectors is unchanged, but the number of ways to achieve them is changed) can impact norms. To see how, suppose that there are two players, $C = \{c_1, c_2, c_3\}$, and $u(c_1) = (0,0)$, $u(c_2) = u(c_3) = (1,1)$. Thus, two consequences out of three lead to payoffs $(1,1)$, and one to $(0,0)$. The aggregate dissatisfaction of c_1 is $D(c_1) = 2 \cdot 2 = 4$: it produces dissatisfaction twice for each player for being worse than $(1,1)$. Now suppose that $C = \{c_1, ..., c_{101}\}$, $u(c_1) = (0,0)$, and all other consequences lead to $(1,1)$. In this case the aggregate dissatisfaction of c_1 is $D(c_1) = 2 \cdot 100 = 200$, since now each player is dissatisfied with c_1 because of 100 other consequences. Intuitively, a bad payoff vector, $(0,0)$, feels much less appropriate if there are more other consequences that lead to a good one. To give an example, suppose you are on a beach and you see someone drowning. In the first case you cannot swim and do not have a phone, but can attract the attention of others (you need to run somewhere). In the second case, you can swim and have a phone, and can attract the attention of others. Intuitively, not helping the drowning person seems less appropriate in the latter case than in the former. Notice that, in order to distinguish situations like this, all consequences should be taken into account like in our definition, and not only those with special properties as in the case of maximal dissatisfaction suggested at the end of Section 2.1.

If we apply our definition of a norm function to the cases with 3 and 101 consequences, we will actually obtain the same result since the norm function is the normalized aggregate dissatisfaction. However, the two cases are connected. In particular, the set of consequences of the first case is a subset of the consequences of the second. We propose a way to compare norm functions in related environments like these in Appendix C. □

Example 5. Ultimatum Game (UG) with punishment norm. From the perspective of our framework UG can be viewed as a DG with a rather extreme punishment mechanism, which only allows to punish a proposer to a maximal degree at the expense of all payoffs in the game.[1] The set of consequences 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 consequence p_c represents rejection choice in the subgame that follows the choice of c.

The left graph on Figure 13 shows the norm function in the UG when considered as a whole. The black line corresponds to the accepted divisions $(1-c, c)$ and the red line stands for the rejection consequences p_c. The right graph shows for each choice $c \in [0,1]$ of a proposer the norm function from the left graph combined with the punishment function μ which is defined, as described in Section 2.2, to be $\mu(c) = -1$ (since the payoff $1-c$ for proposer is the highest attainable in the subgame following offer c) and $\mu(p_c) = 1$ for all $c \neq \frac{1}{2}$ (since p_c is the consequence with the lowest possible payoff for proposer). Notice that for any deviation from equal split the normative valence of consequence p_c is higher than the normative valence of accepting the division $(1-c, c)$. The difference increases for more unequal divisions. □

[1] The Ultimatum Game, first proposed by Güth et al. (1982), was intended as a model of bargaining with rejection representing the "no agreement" option. Thus, it should not be surprising that when we see rejection through the lens of norm-following behavior it seems rather uncalled for as a punishment strategy. In order to test our theory properly an adequate punishment mechanism, which covers appropriate range of payoffs, should be introduced instead (see e.g., Fehr and Fischbacher, 2004).

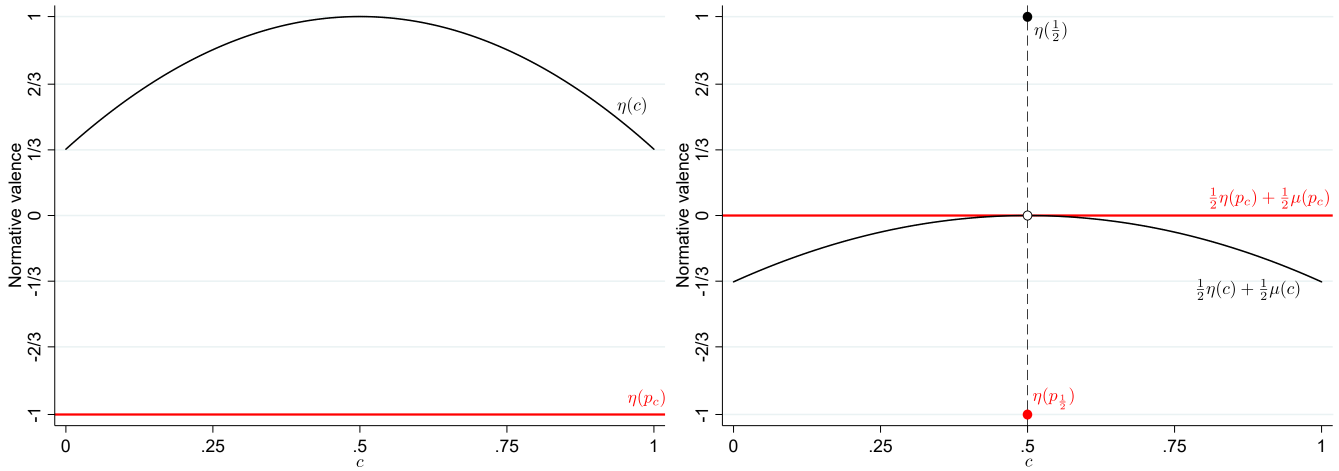
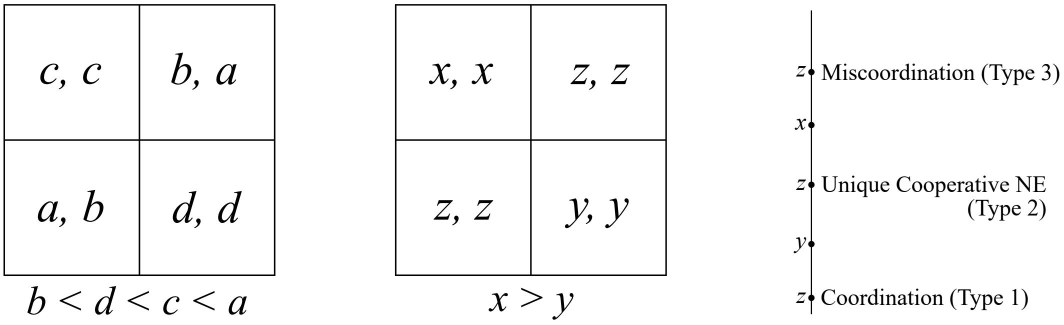


Figure 13: Norm functions in UG. **Left:** the norm function for $\langle N, C, u, R\rangle$. **Right:** for each $c \in C$, except $c = \frac{1}{2}$ which is the norm, convex combination of punishment function and norm function.

Example 6. Prisoner's Dilemma. Consider the Prisoner's Dilemma with material payoffs a, b, c, d as shown on the left graph of Figure 14. We calculate the normative valences associated with each outcome as $x = -2(a-c)$, $y = -4(c-d) - 2(a-c)$, and $z = -3(d-b) - 2(c-d) - (a-c)$.



c, c	b, a
a, b	d, d

$b < d < c < a$

x, x	z, z
z, z	y, y

$x > y$

z Miscoordination (Type 3)
x
z Unique Cooperative NE (Type 2)
y
z Coordination (Type 1)

Figure 14: Prisoner's Dilemma. **Left:** payoffs. **Middle:** normative valences; **Right:** three types of PD that depend on the relationship between normative valences.

Suppose that the two players are extremely rule-following individuals, so that they just want to maximize social appropriateness. Then, the game they play is shown in the middle graph of Figure 14. Depending on the value of z, this game can be of three types: 1) coordination game; 2) dominance solvable with unique NE in which both players cooperate or 3) a miscoordination game (right graph on Figure 14). For the PD of type 1 we obtain *conditional cooperation* behavior: norm abiding players cooperate only if they believe that the other player will cooperate with high enough probability and they defect in the opposite case. Since norm-followers can optimally choose defection or cooperation depending on their beliefs, the observed actions in this kind of PD do not reveal the rule-following propensity of the player. The PD of type 2 is the most clear case where the norm-following players should unambiguously choose cooperation, which also reveals their type. Finally, in the PD of type 3 we may expect mixed strategies and noisy behavior. Thus, our model makes some very specific predictions: cooperation should be the easiest to attain in the type 2 PD, whereas cooperation and defection may coexist in the PD of type 1. □

Case 11. **List (2007).** In Case 3 we analyzed the treatments of this experiment that introduced different giving and taking options to the Dictator game. Here we look at the Take5 treatment and the Earnings treatment, which is the same as Take5 with the only difference being that subjects earned the money that they later were asked to share. In the Take5 treatment all subjects are endowed with $5. The randomly assigned dictators can give some of their endowment to the recipient or take some part of the recipients' endowment (up to full amount of $5). The Earnings treatment is the same as Take5 except subjects earn their endowments by performing a tedious task (sorting and handling charity mail), which induces an ownership claim.

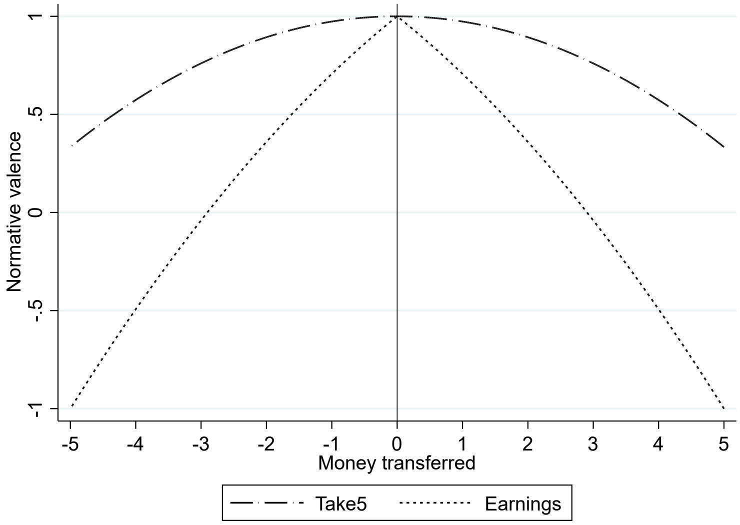


Figure 15: Relative norm functions in the Take5 and Earnings treatments of List (2007).

Figure 15 shows the relative norm functions in the Take5 and Earnings treatments. We model the Earnings treatment by assuming that both dictators and recipients have ownership claims to the $5 that they earned, with weights $\pi_d = \pi_r = 1$ (two ownership classes). As a result, both taking and giving become much less appropriate than in the Take5 treatment with windfall resources. Notice also that the most socially appropriate consequence is to give $0 in both treatments. This is reflected in behavior (Figure 16 on the next page). In the Take5 treatment 30% of subjects choose to give $0 and around 40% to take all money from the recipient. In the Earnings treatment the proportion of subjects who give $0 increases to almost 70%, while the proportion of subjects who take everything from the recipient drops to 20%. These findings are consistent with norm-dependent utility maximization under the norm functions shown in Figure 15, if we assume heterogeneity in the rule-following parameter ϕ_i. Under the model, the treatment effect arises because subjects with medium-low rule-following propensity switch from taking everything in the Take5 treatment to the most socially appropriate option in the Earnings treatment, as the difference between the normative valence of giving $0 and the valence of taking $5 is much higher than in the Take5 treatment. □

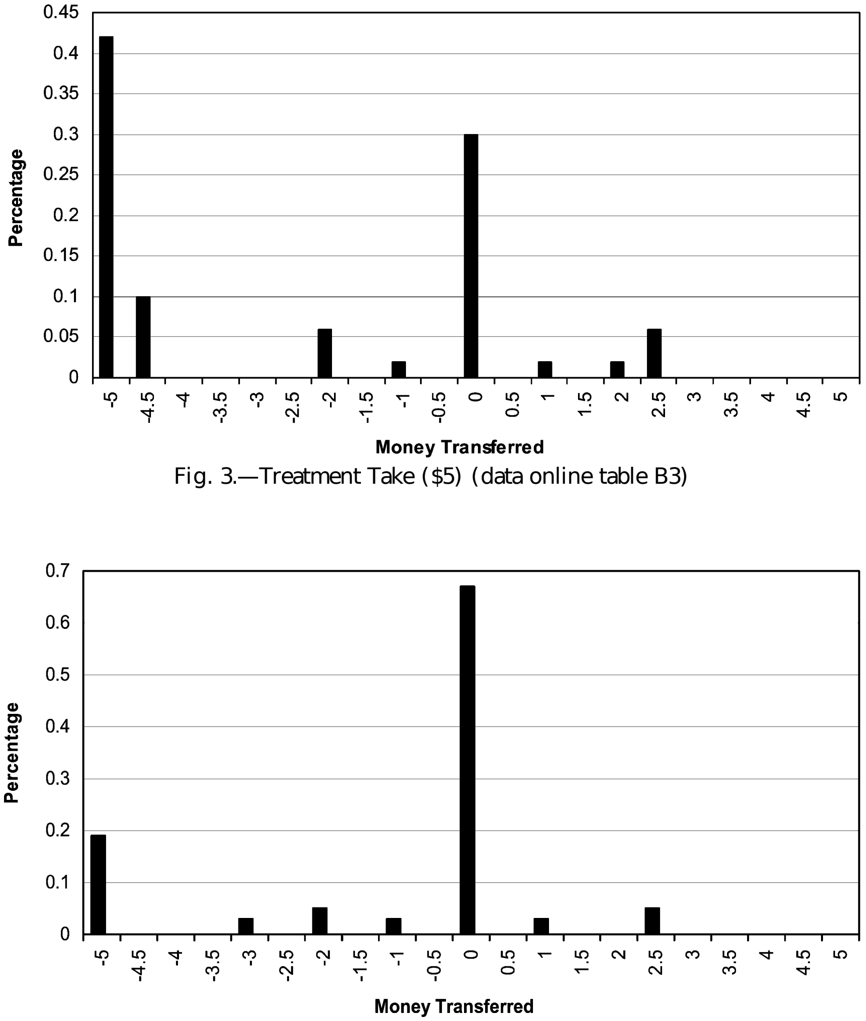


Fig. 3.—Treatment Take ($5) (data online table B3)

Fig. 4.—Treatment earnings (data online table B4)

Figure 16: Data from List (2007).

Case 12. 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). The application of our model to the PD provided in Example 6 in Appendix A 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, 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 of 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 the observations from FF that are hard to interpret with other models. □

A.1 Supporting Evidence for Case 3. List (2007).

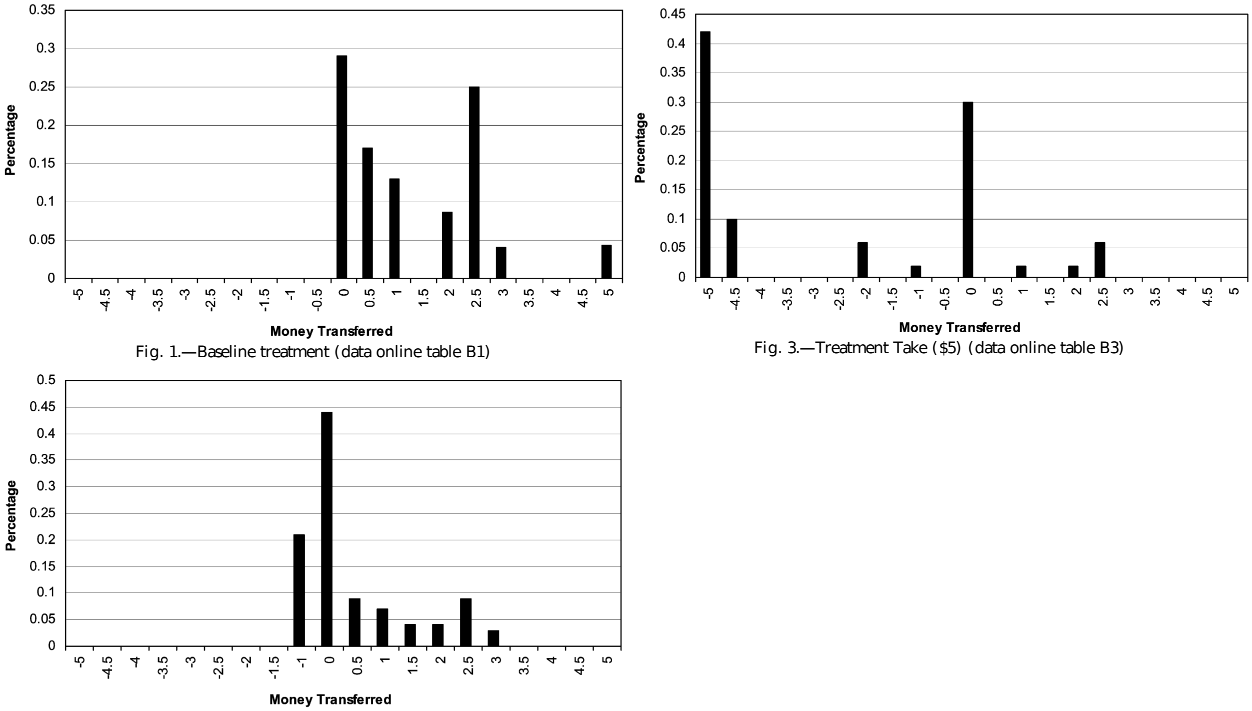


Fig. 1.—Baseline treatment (data online table B1)

Fig. 3.—Treatment Take ($5) (data online table B3)

Fig. 2.—Treatment Take ($1) (data online table B2)

Figure 17: Data from List (2007).

A.2 Supporting Evidence for Case 5. McCabe et al. (2003).

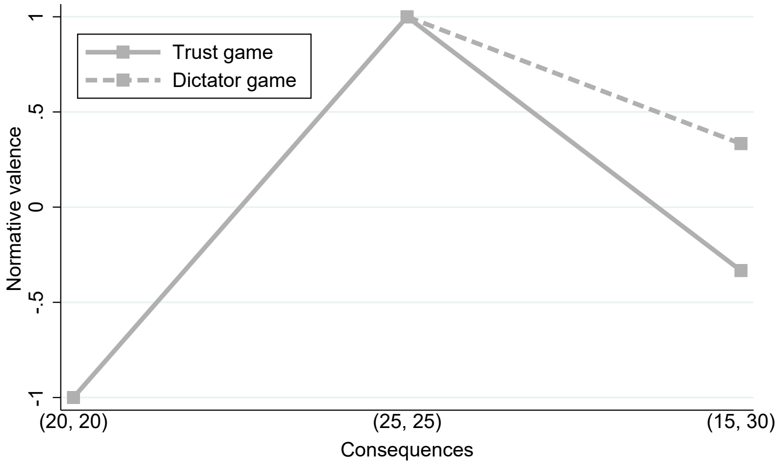


Figure 18: Norm functions with linear utility for the games analyzed in McCabe *et al.* (2003).

A.3 Supporting Evidence for Case 8. Oxoby and Spraggon (2008).

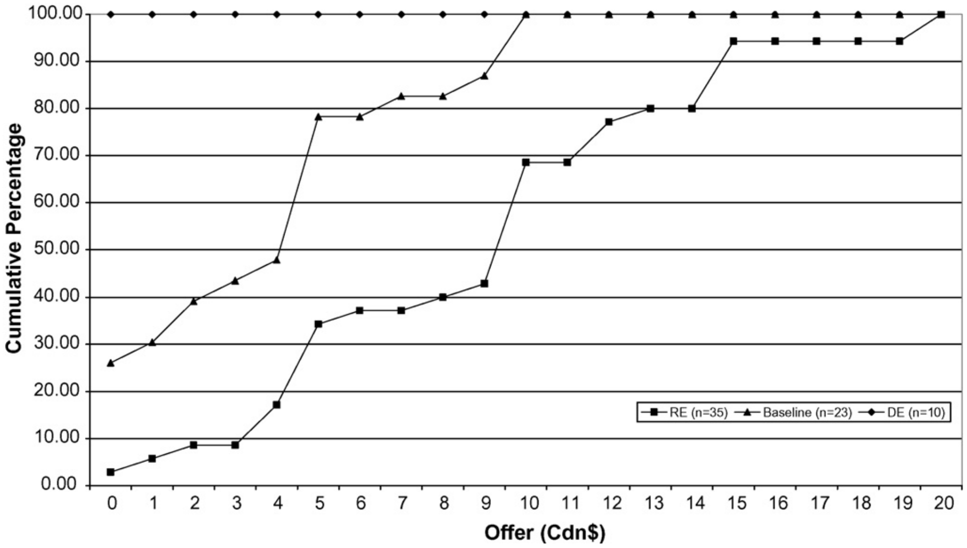


Figure 19: Data from Oxoby and Spraggon (2008).

A.4 Supporting Evidence for Case 10. Chen and Li (2009).

					Choice				Norm Functions										
Game	Role	Payoffs			Ingr.		Outgr.		Ingroup			Outgroup							
		Out	L	R	L	R	L	R		L	R					L	R	ΔChoice	ΔNorms
Dict1	A		400	750	0.30	0.70	0.45	0.55		–1.00	1.00					0.00	1.00	0.15	1.00
	B		400	400															
Dict2	A		400	750	0.67	0.33	0.73	0.27		–1.00	1.00					0.11	1.00	0.06	1.11
	B		400	375															
Dict3	A		300	700	0.68	0.32	0.86	0.14		–1.00	1.00					0.27	1.00	0.18	1.27
	B		600	500															
Dict4	A		200	600	0.34	0.66	0.63	0.38		–1.00	1.00					0.16	1.00	0.29	1.16
	B		700	600															
Dict5	A		0	400	0.56	0.44	0.77	0.23		–1.00	1.00					0.07	1.00	0.21	1.07
	B		800	400															
									Ingroup			A's norm f.			B's norm f.				
									Out	L	R	Out	L	R	Out	L	R		
Resp1a	A	750	400	750	0.26	0.74	0.48	0.53	–1.00	0.88	1.00	0.00	0.88	1.00	–1.00	0.94	1.00	0.22	0.06
	B	0	400	400															
Resp1b	A	550	400	750	0.39	0.61	0.55	0.45	1.00	–1.00	0.98	0.68	–0.98	1.00	1.00	–0.33	0.66	0.16	0.99
	B	550	400	400															
Resp6	A	100	75	125	0.18	0.83	0.33	0.68	1.00	–1.00	–0.40	1.00	–0.22	0.38	1.00	–0.78	–0.48	0.15	0.30
	B	1000	125	125															
Resp7	A	450	200	400	0.10	0.90	0.29	0.71	1.00	–1.00	0.20	1.00	–0.65	0.55	1.00	–0.35	0.25	0.19	0.60
	B	900	400	400															
Resp2a	A	750	400	750	0.67	0.33	0.80	0.20	–1.00	0.89	1.00	0.00	0.88	1.00	–1.00	0.95	1.00	0.13	0.06
	B	0	400	375															
Resp2b	A	550	400	750	0.68	0.32	0.84	0.16	1.00	–1.00	0.71	0.82	–0.85	1.00	1.00	–0.33	0.39	0.16	0.99
	B	550	400	375															
Resp3	A	750	300	700	0.56	0.44	0.73	0.27	–0.98	0.05	1.00	0.03	–0.01	1.00	–1.00	0.58	1.00	0.17	0.53
	B	100	600	500															
Resp4	A	700	200	600	0.35	0.65	0.58	0.42	–0.93	–0.93	1.00	0.11	–1.00	1.00	–1.00	0.11	1.00	0.23	1.04
	B	200	700	600															
Resp5a	A	800	0	400	0.46	0.54	0.59	0.41	–0.97	–0.97	1.00	0.05	–1.00	1.00	–1.00	0.05	1.00	0.13	1.02
	B	0	800	400															
Resp5b	A	0	0	400	0.54	0.46	0.76	0.24	–0.86	–0.86	1.00	–1.00	–1.00	1.00	0.21	0.21	1.00	0.22	1.07
	B	800	800	400															
Resp8	A	725	400	750	0.66	0.34	0.76	0.24	–1.00	0.89	1.00	0.00	0.89	1.00	–1.00	0.95	1.00	0.10	0.06
	B	0	400	375															
Resp9	A	450	350	450	0.69	0.31	0.78	0.23	–1.00	0.98	1.00	0.00	0.96	1.00	–1.00	1.00	1.00	0.09	0.02
	B	0	450	350															
Resp10	A	375	400	350	0.99	0.01	0.96	0.04	1.00	–0.29	–1.00	1.00	0.40	–0.10	1.00	–0.34	–0.90	–0.03	–0.15
	B	1000	400	350															
Resp11	A	400	400	0	0.95	0.05	0.89	0.11	1.00	0.92	–1.00	1.00	0.96	–0.50	1.00	0.92	–0.50	–0.06	–0.50
	B	1200	200	0															
Resp12	A	375	400	250	0.93	0.08	0.80	0.20	1.00	0.15	–1.00	1.00	0.61	–0.41	1.00	0.11	–0.59	–0.13	–0.44
	B	1000	400	350															
Resp13a	A	750	800	0	0.95	0.05	0.86	0.14	1.00	0.94	–1.00	1.00	0.97	–0.52	1.00	0.94	–0.48	–0.09	–0.52
	B	750	200	0															
Resp13b	A	750	800	0	0.90	0.10	0.84	0.16	1.00	0.91	–1.00	1.00	0.96	–0.85	1.00	0.90	–0.15	–0.06	–0.85
	B	750	200	50															
Resp13c	A	750	800	0	0.91	0.09	0.73	0.28	1.00	0.90	–1.00	1.00	0.95	–0.89	1.00	0.90	–0.11	–0.18	–0.89
	B	750	200	100															
Resp13d	A	750	800	0	0.81	0.19	0.68	0.33	1.00	0.90	–1.00	1.00	0.95	–0.92	1.00	0.89	–0.08	–0.13	–0.92
	B	750	200	150															

Table 2: Normative valences and choices in Chen and Li (2009). ΔChoice is the difference between the Choice L in different- and same-group games. ΔNorms is the difference of differences between B's normative valences of L and R and Ingroup normative valences L and R. In the games player A can first choose Out, which ends the game, or pass the move to player B who chooses between L and R.

B Maximin Preferences

As we show in Section 2.1 our concept of normative valence incorporates both efficiency and equality preferences. However, maximin preferences (choose an allocation with the maximal minimal payoff) are known to have significant explanatory power in many contexts (Engelmann and Strobel, 2004; Baader and Vostroknutov, 2017). In philosophical debates, maximin preferences (Rawls, 1971) are usually counterposed to utilitarianism or maximization of efficiency (Bentham, 1781). However, as we show in this section, the two principles do not have to be considered as different, but can be derived from the single idea that normative appropriateness comes from dissatisfaction.

We propose that the source of differences between efficiency and maximin lies in calculations of dissatisfaction, or utility differences between consequences. Suppose that a player chooses between two allocations for N players as in Example 1. In this example, we implicitly assume that the utility of money is linear and is the same for all players in the game, or, in other words, $u(x) = x$, where x is a monetary payoff in the game. When this is the case, the consequence with the highest efficiency (sum of payoffs) is more appropriate than the less efficient one, and if the two consequences have the same efficiency then their appropriateness is also equal. However, this relationship breaks down if marginal utility of money is decreasing. Put differently, if one player is very rich and another is poor, then taking some amount of money x from the rich will create much less dissatisfaction than the same amount x taken from the poor. This asymmetry will be reflected in the appropriateness of the consequences, with poor player being normatively favored in the same way as it was shown at the end of Example 2. Thus, we see that the more concave the utility of money is the more maximin the preferences expressed by the norm function will become, since the poor players' dissatisfactions will be given relatively more and more weight as compared to the rich players.

Case 13. Millionaires playing DG. Interestingly, there is a perfect evidence of this effect in the literature. Smeets et al. (2015) report the results of a Dictator game played by millionaires and low-income participants. Millionaires, when proposed to split €100 between themselves and a low-income recipient, give on average 71.4% to the recipient, with 45.6% of them giving the entire €100. This is in stark contrast with the average giving of 28.4% in standard Dictator games (Engel, 2011). At the same time, millionaires matched with other millionaires give on average 50%. Such behavior is ideally explained in our framework: the dissatisfaction that millionaires feel from losing €100 is incomparably smaller than the dissatisfaction that low-income person feels when losing the same amount. The dissatisfactions of two millionaires, however, are equal, which results in equal split of money. Thus, the norm favors giving money to the low-income recipient, while prescribing equal split in case of two players with the same income. □

The behavior of millionaires in this case is consistent with the idea of maximin preferences: maximize the payoff of the poorest player. This case suggests that maximin preferences manifest themselves when there is a noticeable difference in income levels of the players. In "standard" experiments with university students this effect can also be spotted, though it is, understandably, not as pronounced as in Case 13 above. In experiments reported in Engelmann and Strobel (2004) and Baader and Vostroknutov (2017), subjects choose allocations for three players. In both studies, by circumstance, one of the three players was assigned much smaller payoff than the other two. This created a similar situation when the dissatisfaction of this "poor" player were seemingly more important than those of other players. As a result, in all mini-tasks where maximin and efficiency favored different outcomes around half of the subjects chose maximin and another half efficiency, while in cases when efficiency and maximin were aligned the choices of almost all subjects were aligned as well (see Case 1 in Section 3.1). This demonstrates that subjects differ in how they assess the normative valences of consequences. While some use monetary payoffs to make inferences about which consequence is more appropriate, some others take into account the relative income of the players, as if performing a concave transformation of payoffs. Interestingly, Baader and Vostroknutov (2017) report that students with economics background favor efficiency principle, while students without such background (Arts and Culture, European Studies) favor maximin. This finding is in line with this idea: economics students are taught to perceive allocations in terms of monetary gains and losses, while non-economics students are more familiar with general thinking related to poverty and inequality.

To model such situations we propose a simple change in the calculation of dissatisfaction:

$$d_i(x, c) = \max\{f(u_i(c)) - f(u_i(x)), 0\}, \tag{7}$$

Here $u_i(x)$ is thought of as a monetary payoff of player i in consequence x and f is a concave increasing function that represents diminishing marginal utility of money. Thus, for a fixed difference between two low payoffs the dissatisfaction is higher than for the same difference between high payoffs. We demonstrate how the concept works with two examples.

Example 7. Suppose the set of consequences is $C = \{a, b\}$ with two players 1 and 2, and the payoffs are defined as $u(a) = (I_1, I_2 + x)$, $u(b) = (I_1 + x, I_2)$. In words, players have incomes I_1 and I_2 and one of them chooses whether an amount x goes to first or second player. Notice that the efficiency of the two allocations, in terms of money, is the same. The aggregate dissatisfaction of a is $D(a) = f(I_1 + x) - f(I_1)$ and the aggregate dissatisfaction of b is $D(b) = f(I_2 + x) - f(I_2)$. Then if $I_1 > I_2$, we have $D(a) < D(b)$. This means that consequence a, where x goes to the poorer player, is more appropriate than the one in which x goes to the richer player. This is consistent with maximizing minimal payoff. □

Example 8. Consider a DG with consequences $C = [0, 1]$ and utilities defined by $u(c) = (I + 1 - c, 5 + c)$. Here I is the income of the dictator and 5 is the income of the recipient. Suppose that we transform the payoffs with the function $f(x) = \ln x$.

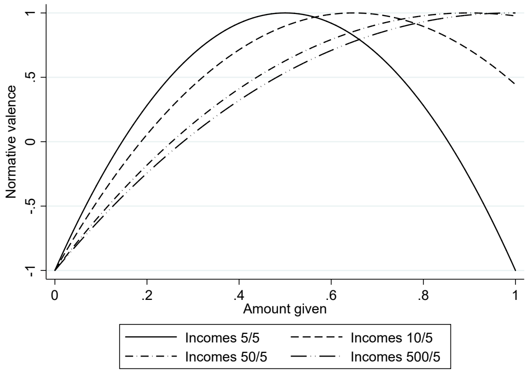


Figure 20: Norm functions in the DG with income of dictator changing from 5, to 5, to 500.

Figure 20 shows the norm functions for $I = 5, 10, 50, 500$. As income of the dictator grows relatively to that of the recipient, the norm increases from $c^* = 0.5$ to 0.65 to 0.9 to 0.99. Without transforming the payoffs with f, the norm is always $c^* = 0.5$. Thus, a concave transformation is necessary in our model to explain maximin behavior. □

C Comparison of Norm Functions across Environments

In this appendix we discuss how to compare norm functions between environments. This is mostly important when the norm functions in different treatments of the same experiment or in otherwise related situations should be compared. We use Example 4 to provide intuition. In the example a bystander who sees someone drowning has options “do nothing” and “attention” in one environment, but in a related environment he can also “call” and “swim.” The consequence “do nothing” is intuitively less appropriate when one has more opportunities to help a drowning stranger.

We need a way to compare normative valences of a consequence c that belongs to two different sets of consequences C_1 and C_2 ($c \in C_1 \cap C_2$). We assume that the payoffs from c, $u(c)$, are the same in both sets of consequences. We treat the environments $\langle N, C_1, u_1, D^1 \rangle$ and $\langle N, C_2, u_2, D^2 \rangle$ as separate and possessing their own norm functions η_{C_1} and η_{C_2}. In order to compare these norm functions on $C_1 \cap C_2$, we need to find some common ground, since η_{C_1} and η_{C_2} are normalized using completely different dissatisfactions. We postulate that if a consequence $c_i \in C_i$, $i \in \{1, 2\}$, is the most appropriate in its corresponding set or $\eta_{C_i}(c_i) = 1$, then it should also be the most appropriate in the *relative norm function* that we construct below. In other words, the appropriateness of the best consequence does not depend on the relative comparisons made. The normative valences for all other consequences are normalized using dissatisfactions in *both* C_1 and C_2. In particular, let $m_i = \min_{c \in C_i} D^i(c)$ and $m = \min_{i \in \{1,2\}} m_i$, and redefine the dissatisfactions as $\bar{D}^i(c) = D^i(c) - m_i + m$, so that the lowest dissatisfaction (for the most appropriate consequence) is the same in both environments.[2] Let $x = \max_{i \in \{1,2\}} \max_{c \in C_i} \bar{D}^i(c)$ be the highest dissatisfaction in all environments and use the interval $[m, x]$ for normalization of all aggregate dissatisfaction functions.

Definition 2. *For $\langle N, C_1, u_1, D^1 \rangle$ and $\langle N, C_2, u_2, D^2 \rangle$, call $\ddot{\eta}_{C_i} : C_i \to [-1, 1]$ defined as $\ddot{\eta}_{C_i}(c) := [-\bar{D}^i(c)]_{[-x,-m]}$ a **relative norm function** or **norm function relative to** C_{-i}.*

In this definition, first, the aggregate dissatisfactions D^1 and D^2 are computed and then the relative norm function $\ddot{\eta}_{C_1}$ is calculated as $-\bar{D}^1$, which is normalized from the interval that covers dissatisfactions in both environments to $[-1, 1]$.

Example 9. Drowning example with relative norm functions. We return to the two situations presented in Example 4. Recall that $C = \{\text{do nothing}, \text{swim}, \text{call}, \text{attention}\}$ and $C_1 = \{\text{do nothing}, \text{attention}\}$ with utilities $u(\text{do nothing}) = (0, 0)$ and $u(c) = (1, 1)$ for all other consequences c. For the superset C we have $D(\text{do nothing}) = 3$ and $D(c) = 0$ for other consequences. For C_1 we have $D^1(\text{do nothing}) = 1$ and $D^1(\text{attention}) = 0$ for other consequences. Thus, $\eta_C(\text{do nothing}) = -1$ and $\eta_C(c) = 1$ for the other consequences. From Definition 2 we obtain $\ddot{\eta}_{C_1}(\text{do nothing}) = \frac{1}{3}$ and $\ddot{\eta}_{C_1}(\text{attention}) = 1$. Thus, the appropriateness of doing nothing is higher when there are few options to help, exactly as our intuition had it. □

[2] Note that adding a constant to D^i or multiplying it by a positive constant does not change the associated norm function η_{C_i}.

D Norm-Dependent Utility in Games with Observable Actions

Let a tuple $\Gamma = \langle N, C, u, D, H \rangle$ be an *extensive form game with observable actions*, where $\langle N, C, u, D \rangle$ is an environment with the set of consequences C corresponding to the set of terminal nodes and H is 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 $a^t = (a^t_1, ..., a^t_N)$ 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 $\{\emptyset\} \in H$ 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.

D.1 Games without a Separate Punishment Mechanism

We start with the setup without specially defined punishment mechanisms. 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 consequences 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, $\eta^{\{\emptyset\}} = \eta_C$.

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 $\mu^{a_i}_i$ 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 $\mu^{a_i}_i$ and the way they are combined with $\eta^{p(h)}$.

To define $\mu^{a_i}_i$ we use the same logic as in Section 2.2. We determine the degree of norm violation of each player i and construct the punishment in the i's payoff space. Let $C^{a_i}_{p(h)} \subseteq C_{p(h)}$ be the set of consequences reachable given the choice a_i of player i. Notice that $C^{a_i}_{p(h)}$ is weakly larger than C_h, the set of consequences 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 consequences that should be used for the determination of punishment since player i cannot be held responsible for what other players choose. Let $M_{p(h)} = \text{argmax}_{c \in C_{p(h)}} \eta^{p(h)}(c)$ be the set of the most appropriate consequences according to $\eta^{p(h)}$. In the simplest case, all players choose actions a_i that leave some consequences 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

$$\eta^h = \eta^{p(h)} \quad \text{if } \forall_{i \in N} \;\; C^{a_i}_{p(h)} \cap M_{p(h)} \neq \emptyset.$$[3]

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 consequences 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

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

[3] Notice that this definition allows for the possibility that each player chooses the action consistent with some consequence in $M_{p(h)}$, but the resulting action profile $a = (a_i)_{i \in N}$ makes all consequences in $M_{p(h)}$ unreachable. This, for example, happens in Type 3 Prisoner's Dilemma described in Example 6 when both players choose to defect. We take the stance that players should not be punished in this circumstance, but alternative definitions can as well be considered.

$r_i^{a_i}$ is positive only for players who chose the actions inconsistent with all consequences in $M_{p(h)}$. Let $V = \{i \mid r_i^{a_i} > 0\} \subseteq N$ 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 consequence, $u_{im} = \max_{c \in M_{p(h)}} u_i(c)$, or the payoff that she chose to forgo when choosing a_i; 2) the minimal payoff that she can obtain in the whole game, $\underline{u}_i = \min_{c \in C} u_i(c)$, which serves as a reference point for the harshest punishment;[4] and 3) the payoff that i "aims at" by choosing a_i, $\bar{u}_i = \max_{c \in C^{a_i}_{p(h)}} u_i(c)$. Let $m_i = \min\{u_{im}, \bar{u}_i\}$ and define the punishment norm function $\mu_i^{a_i}$ as shown in Figure 1 in Section 2.2. Finally we calculate the norm function η^h by combining $\eta^{p(h)}$ and the punishment functions $(\mu_i^{a_i})_{i \in V}$:

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

where $\mu_i^{a_i}(c)$ is short for $\mu_i^{a_i}(u(c))$. Essentially, η^h is a convex combination of $\eta^{p(h)}$ and the average punishment function that gives equal weights to all players.[5]

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 $\eta^{\{\emptyset\}} = \eta_C$, we can 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 $\Gamma' = \langle N, C, w, H \rangle$ be the same game only with utilities defined by

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

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

D.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.[6]

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

[5] 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 $\max\{\mu_1^{a_1}(c), ..., \mu_N^{a_N}(c)\}$. In Section 4.2 we also propose that players may have their personal punishment weights that reflect their role entitlements, in which case the average should be replaced with a weighted sum.

[6] 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.

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_C 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_C 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 $\eta^{p(h)} = \eta_C$ on its domain. $\mu_i^{a_i}$ 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; $\sigma + (1-\sigma)\mu_i^{a_i}(s)$ is the punishment norm function adjusted with the weight σ as in Section 2.2; and $\zeta(x)$ is an increasing cost function with $\zeta(0) = 0$.[7] $s_{ji}^{a_i}$ 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.[8]

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 $\Gamma'' = \langle N, C, v, H \rangle$ with utility for player i calculated as follows. For any consequence c, which is also a terminal history, let us write $c = (a^1, a^2, ..., a^\ell)$, where $a^t = (a_1^t, ..., a_N^t)$ is the action profile chosen in stage t that leads to c. Let

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

The utility v_j is simply the norm-dependent utility with the norm function η_C minus the punishment that player j incurs on the way to c and the cost of punishment that j metes upon 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. We do not claim that this is the way it should be modeled, but just propose one possibility how it can be done.

D.3 Games in Context

When injunctive norm functions depend on social parameters like status, ingroup, kin, or ownership claims, we have different norm functions for different players, say η_C^j for player j. In this case, the same derivations of norm-dependent utilities as in the two sections above are possible for each player separately starting from the norm function η_C^j at the beginning of the game. Thus, to compute norm-dependent utilities for a game in context one needs to repeat the derivations above for each player and to modify only that player's payoffs.

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

[8] By the definition of μ_i in Section 2.2, 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.

E Proofs

Proof of Proposition 1. Consider an environment $\langle N, C, u, R\rangle$ and two consequences $c_1, c_2 \in C$ with $u_i(c_1) \geq u_i(c_2)$ for all $i \in N$ with at least one strict inequality. For any i with $u_i(c_1) = u_i(c_2)$ we have $D_i(c_1) = D_i(c_2)$, and for any i with $u_i(c_1) > u_i(c_2)$ it is true that

$$D_i(c_1) = \int_{c \in C} \max\{u_i(c) - u_i(c_1), 0\} dc \leq \int_{c \in C} \max\{u_i(c) - u_i(c_2), 0\} dc = D_i(c_2).$$

Thus, $D(c_1) \leq D(c_2)$. When C is finite the inequality is strict since $d_i(c_2, c_1) > 0$ for i with $u_i(c_1) > u_i(c_2)$. When C has cardinality of the continuum, it is possible to have $D(c_1) = D(c_2)$. For example, for two players suppose that the image $u[C] = \{(x,x) | x \in [0,1]\} \cup (2,2)$. The payoffs in two consequences are $u(c_1) = (2,2)$ and $u(c_2) = (1,1)$. Then c_1 Pareto dominates c_2, but $D(c_1) = D(c_2) = 0$. ■

Proof of Proposition 2. For any consequence c_j the aggregate dissatisfaction is given by

$$D(c_j) = \sum_{i=1}^{j-1} (u_j - u_i) + \sum_{i=j+1}^{K} (u_i - u_j),$$

which can be rewritten as

$$D(c_j) = \sum_{i=1}^{j-1} i(u_{i+1} - u_i) + \sum_{i=j+1}^{K} (K - i + 1)(u_i - u_{i-1}).$$

From this it follows that for all $j = 1..K - 1$

$$D(c_{j+1}) - D(c_j) = (2j - K)(u_{j+1} - u_j).$$

The difference is (weakly) negative for $j < \frac{K}{2}$ and positive for $j > \frac{K}{2}$. Thus, the consequences with the smallest aggregate dissatisfaction are $j = \frac{K}{2}$ and $j = \frac{K}{2} + 1$ if K is even, and $j = \frac{K}{2} + \frac{1}{2}$ is K is odd. ■

Additional References in Appendices

Baader, M. and Vostroknutov, A. (2017). Interaction of reasoning ability and distributional preferences in a social dilemma. *Journal of Economic Behavior & Organization*, 142:79–91.

Bentham, J. (1781). An introduction to the principles of morals and legislation. *History of Economic Thought Books*.

Cappelen, A. W., Nielsen, U. H., Sørensen, E. Ø., Tungodden, B., and Tyran, J.-R. (2013). Give and take in dictator games. *Economics Letters*, 118(2):280–283.

Chen, Y. and Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, 99(1):431–57.

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, 14(4):583–610.

Engelmann, D. and Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution. *American Economic Review*, 94(4):857–869.

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

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

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, 115(3):482–493.

McCabe, K. A., Rigdon, M. L., and Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization*, 52:267–275.

Oxoby, R. J. and Spraggon, J. (2008). Mine and yours: Property rights in dictator games. *Journal of Economic Behavior & Organization*, 65(3-4):703–713.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Smeets, P., Bauer, R., and Gneezy, U. (2015). Giving behavior of millionaires. *Proceedings of the National Academy of Sciences*, 112(34):10641–10644.