

# A Theory of Injunctive Norms\*

Erik O. Kimbrough<sup>†</sup>

Alexander Vostroknutov<sup>‡ §</sup>

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.

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

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

<sup>§</sup>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*

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.<sup>1</sup> 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.<sup>2</sup> 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

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).<sup>3</sup> 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

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 co-equal 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 sparse 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.<sup>5</sup> 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.<sup>6</sup> 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

---

<sup>5</sup>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.

<sup>6</sup>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  $u : C \rightarrow \mathbb{R}^N$  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  $x \in C$ .<sup>7</sup>

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

---

<sup>7</sup>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

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

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.<sup>8</sup>

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

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

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.<sup>9</sup>

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

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

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.<sup>10</sup> 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.

---

<sup>8</sup>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.

<sup>9</sup>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  $x$ .

<sup>10</sup>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.<sup>11</sup> 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  $h_C : C \rightarrow [-1, 1]$ , defined as

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

where  $[ \ ]_{\text{Conv}(D)}$  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  $h_C(x) = 1$  for all  $x \in C$ .

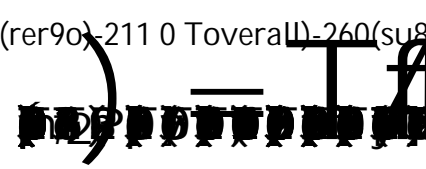
In this definition,  $h_C$  is simply the negative of aggregate dissatisfaction, normalized to the interval  $[-1, 1]$ . Thus, the consequence  $x$  with  $h_C(x) = 1$  is the most socially appropriate (the norm) and the one with  $h_C(x) = -1$ , the least socially appropriate. If all consequences have the same aggregate dissatisfaction, then we assume that  $h_C(x) = 1$  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).<sup>12</sup>

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  $(a_1, \dots, a_N)$  and  $(b_1, \dots, b_N)$  and suppose that  $a_1 + \dots + a_N > b_1 + \dots + b_N$ , or the efficiency of  $c_1$  is higher than the efficiency of  $c_2$ . This inequality can be rewritten as

$$\sum_{i: a_i > b_i} a_i - b_i > \sum_{i: a_i < b_i} b_i - a_i, \quad D(c_2) > D(c_1) \implies h_C(c_1) = 1 \text{ and } h_C(c_2) = 0.$$

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

P(7)-234(ar)18(e)-235(not)-235(rer9o)-211 0 Toverall)-260(su8  


This is an upward sloping parabola which is minimized at  $c = \frac{1}{2}$ . Thus, the norm function  $h_C$  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  $u_r(c) = gc$ , where  $g > 0$ . To illustrate, suppose that receiver is in dire circumstances and his  $g$  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  $g$  included, we find that the norm is now  $c = g/(1 + g)$ , which goes to 1 as  $g$  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  $(N, C, u, D)$  has two players and  $K$  consequences  $c_1, c_2, \dots, c_K$  with utilities  $u_1, u_2, \dots, u_K$  for one player and  $a_1, a_2, \dots, a_K$  for the other ( $a_1, \dots, a_K \geq \mathbf{R}$ ). 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,  $D(x) = u - \frac{1}{N} \sum_{i \in N} u_i(x)$ , where  $u$  is the sum of highest payoffs of all players. This means that  $h_C(x)$  is a positive affine transformation of the sum of payoffs in  $x$ . So, the alternative dissatisfaction integration method ranks consequences

## 2.2 Punishment

that can still be obtained after  $a$  was chosen ( $\max_{c \in C_a} h_C(c)$ ).<sup>14</sup>

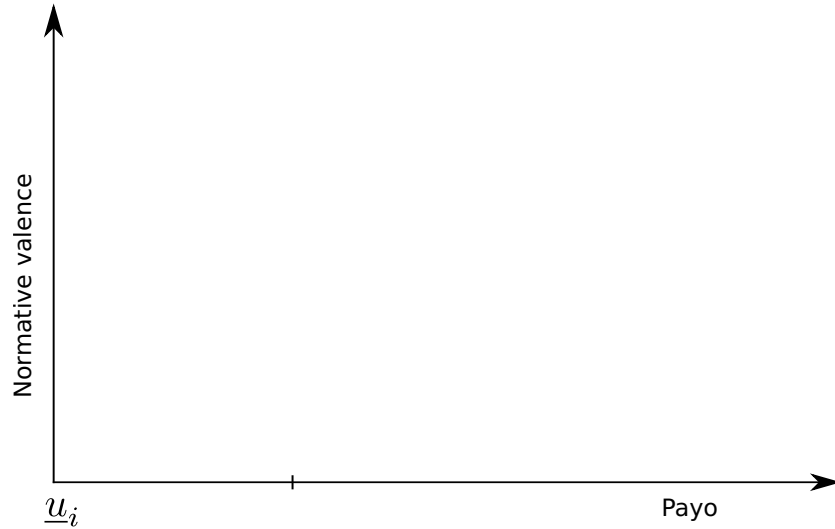


Figure 1: Punishment norms for  $r_a < 2$  (solid gray line),  $r_a = 2$  (dashed line), and  $r_a \neq 0$  (dotted line).

Notice first that  $m_i$  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  $m_i$  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  $\frac{r_a}{2}\underline{u}_i + 1 - \frac{r_a}{2}m$ . This value is linearly proportional to  $r_a$  and is equal to  $m$  when  $r_a \neq 0$  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  $\frac{r_a}{2}\underline{u}_i + 1 - \frac{r_a}{2}m$  have the highest normative valence of 1. This implies that for the punishers it is normatively irrelevant whether  $i$  gets punished by having payoff  $\frac{r_a}{2}\underline{u}_i + 1 - \frac{r_a}{2}m$  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  $s + (1 - s)m_i(p)$  for payoff  $p$  of player  $i$ . The parameter  $s \in [0, 1]$  represents the relative importance of punishment in a given situation. When  $s = 1$  all punishment options are equally (and maximally) socially appropriate; thus, the least costly punishment

will be chosen. When  $s = 0$ , 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  $h_C$  and  $m_i$ , thereby, increasing the normative valence of the consequences that decrease  $i$ 's payoff. Abusing notation, let us think of the function  $m_i$ , originally defined on the space of payoffs, as a function defined on consequences with  $m_i(x)$  meaning  $m_i(u_i(x))$  and assume that for each  $x \in C_a$  the combined norm function is

$$h^\theta(x) = sh_C(x) + (1 - s)m_i(x).$$

Here, again, the parameter  $s$  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.<sup>17</sup>

### 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.<sup>18</sup>

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



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

$$w_i(x) := u_i(x) + f_i h(x),$$

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.  $h(x)$  is the normative valence of  $x$  in the game node directly leading to  $x$  (pre-terminal node) if no separate punishment mechanism is available.<sup>19</sup> When there exists a punishment mechanism,  $h(x)$  is the same as  $h_C(x)$ , the norm function defined by all consequences in the game (see Appendix D).  $f_i \geq 0$  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  $f_i = 0$  we have a standard utility maximizer, as  $f_i \rightarrow \infty$  we have player  $i$  who only cares about following norms.<sup>20</sup>

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

$$h_{a_1}(c) = s h_C(c) + (1 - s) m_1(c|a_1) \text{ for } c \in C_{a_1},$$

as defined in Section 2.2. Here notation  $m_1(c|a_1)$  means the punishment norm function that is calculated for the consequences  $C_{a_1}$ . Norm functions  $h_{a_1}$  are defined on the pre-terminal node

### 3.1 Choice-Set-Dependent Social Preferences

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

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.

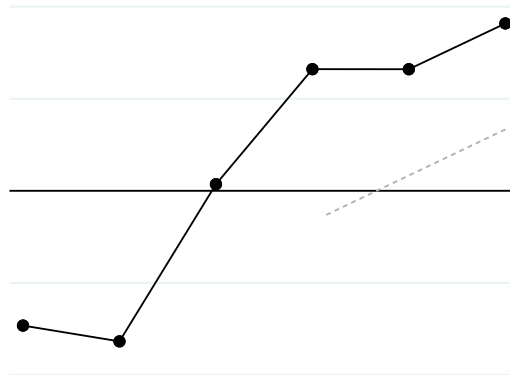


Figure 3: Solid line represents the data from Galeotti et al. (2018). Dashed lines show the predictions implied by norm functions computed under both the linear and the log utility models (grey and black lines respectively).

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, speci-

### 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  $f_i$  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).

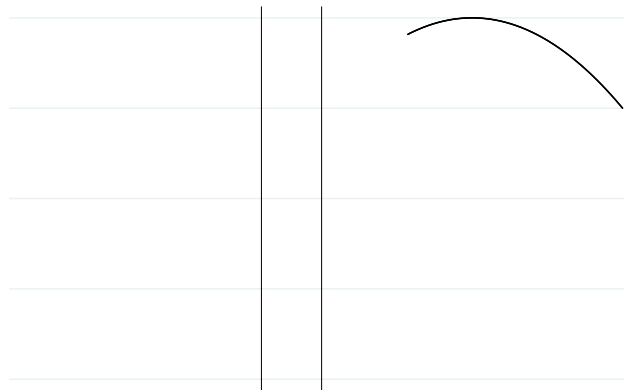


Figure 4: Relative norms in the three treatments of [List \(2007\)](#).

The graph on Figure 4 shows the *relative* norm functions in the Baseline, Take1, and Take5 treatments.<sup>21</sup> 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

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 Kirchsteiger 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  $f_i$  in the utility, see Section 2.3), and some do not (low  $f_i$ ).

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*.

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” (



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 behavior

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)

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.

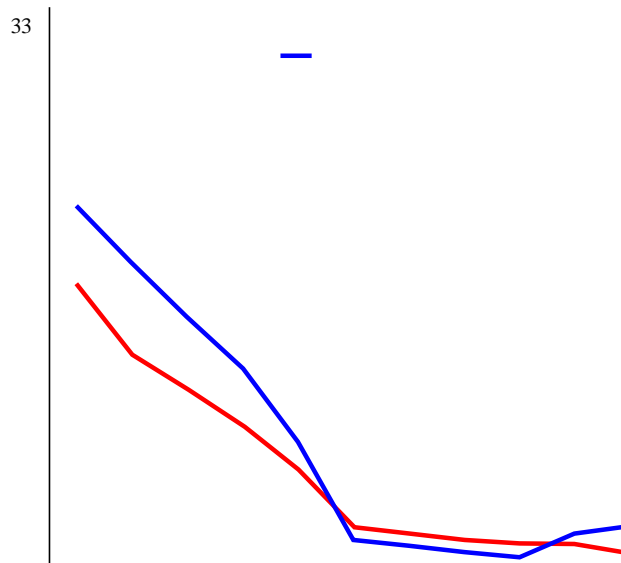


Figure 8: Third and second party punishment in the DG reported in FF. The dashed line shows the model predictions of the harshest possible punishment meted by the extreme norm-followers with very little costs of punishment.

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



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  $u: C \rightarrow \mathbb{R}^{NP}$ , 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  $x \succeq C$ . Finally, let  $p_{ip} \in [0, 1]$  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  $\sum_{i \in N} p_{ip} = 1$ .

For example, suppose that player  $i$  "completely" owns the resources in class  $p$ . Then  $p_{ip} = 1$  and  $p_{jp} = 0$  for all other players  $j \neq i$ . Player  $i$  might also have a weaker ownership claim. In this case we can have  $p_{ip} = 0.8$ ,  $p_{kp} = 0.2$ , and  $p_{jp} = 0$  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  $p^\emptyset$  in which  $\sum_{i \in N} p_{ip^\emptyset} = p_{jp^\emptyset}$ .

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

$$d_i(x, c) := \max_{p \in P} \sum p_{ip} (u_{ip}(c) - u_{ip}(x)), 0. \quad (4)$$

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  $p_p = 1$  and  $p_r = 0$  (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

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

$p$

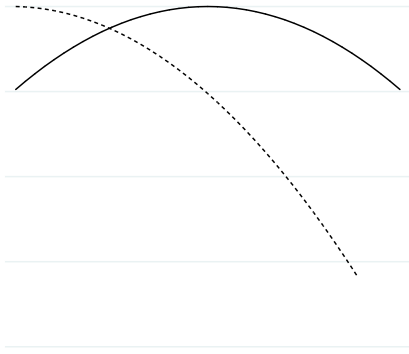


Figure 9: **Left:** relative norm functions in the three treatments of [Oxoby and Spraggon \(2008\)](#): Recipient Earnings (RE), Baseline, and Dictator Earnings (DE). **Right:** cumulative distributions of offers in these treatments.

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 re-





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,  $s_p = 0.53$ ).<sup>26</sup> 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.<sup>27</sup>

### 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

---

<sup>26</sup>To clarify notation, check Example 5 in Appendix A that describes standard UG.

<sup>27</sup>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  $t_i \in [0, 1]$  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  $r_{k'}$   $\in [0, 1]$  where  $k$  is the index of the in-group and  $'$  is the index of the out-group ( $k, ' \in \{1, 2\}$ ). In principle,  $r_{k'}$  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.<sup>28</sup> The aggregate dissatisfaction from the perspective of group  $N_1$  is

$$D(x$$

Finally, for each player  $j \in N$  we define kin relationships by means of weights  $k_{ji} \in [0, 1]$ , where  $i \in N$  indexes all other players. We assume that  $k_{jj} = 1$  for all  $j \in N$  (own dissatisfaction is counted as the most important) and that

task and ask how well those weights (and the implied norms) predict play in a subsequent series of games.

**Case 10.**



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  $p$  that determine ownership claims, punishment weights  $s$  for the role entitlements, and weights  $t, r, k$  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 counter-punishment 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. *aslexpgview*, ex97(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). Rhetor1nE1044(matters:)-306(a)-244(social)-244(norms)-243(ex)gence of human prosociality. *Trends in cognitive sciences*, 15(5):218–226.
- Cox, J. C., List, J. A., P M., Sadiraj, V., and Samek, A. (2018). Moral costs and rational cho-2e: Theory and experimental evidence. mimeo, Georgia State University, University of Ch-25go, 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*, exchapter 20, expages 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). Mixingm(85.e(and)-381(Economic)-380(bee(18(obeIT182onomic)-

- 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*





Figure 13: Norm functions in UG. **Left:** the norm function for  $(hN, C, u, Ri)$ . **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  $z = 3(d - b) - 2(c - d) - (a - c)$ .

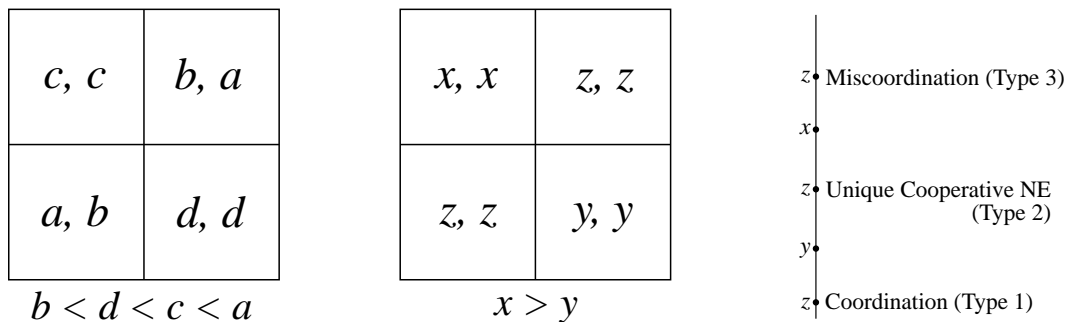


Figure 14: Prisoner's Dilemma. **Left:** payoffs. **Middle:** normative valences; **Right:** three types of PD that depend on the relationship between normative valences.

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  $p_d = p_r = 1$  (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  $f_i$ . 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.

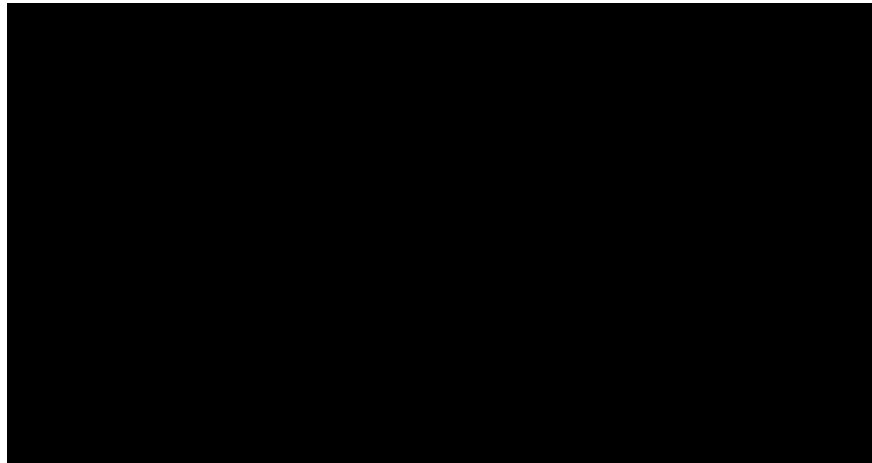
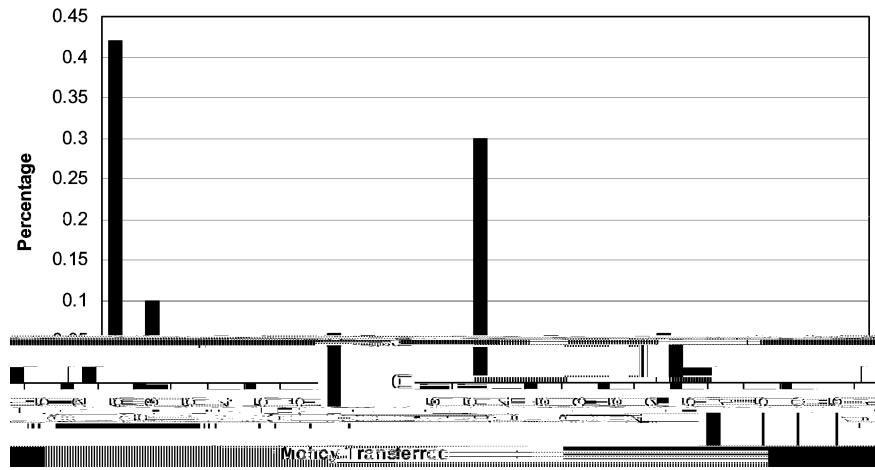


Figure 16: Data from List (2007).

**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

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

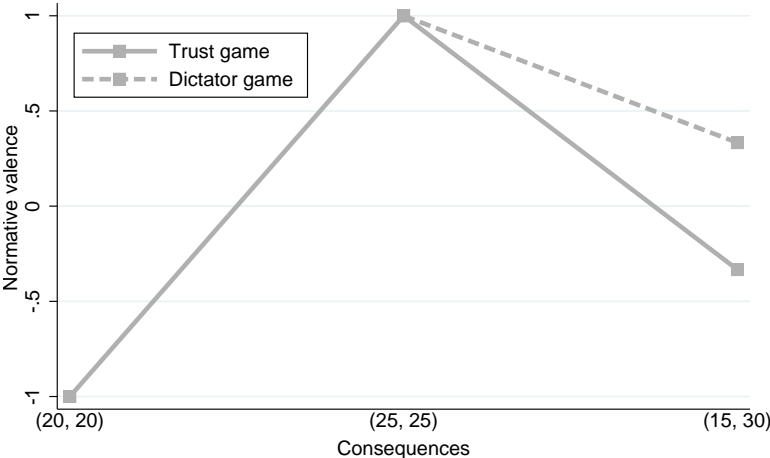


Figure 18: Norm functions with linear utility for the games analyzed in McCabe et al. (2003).

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

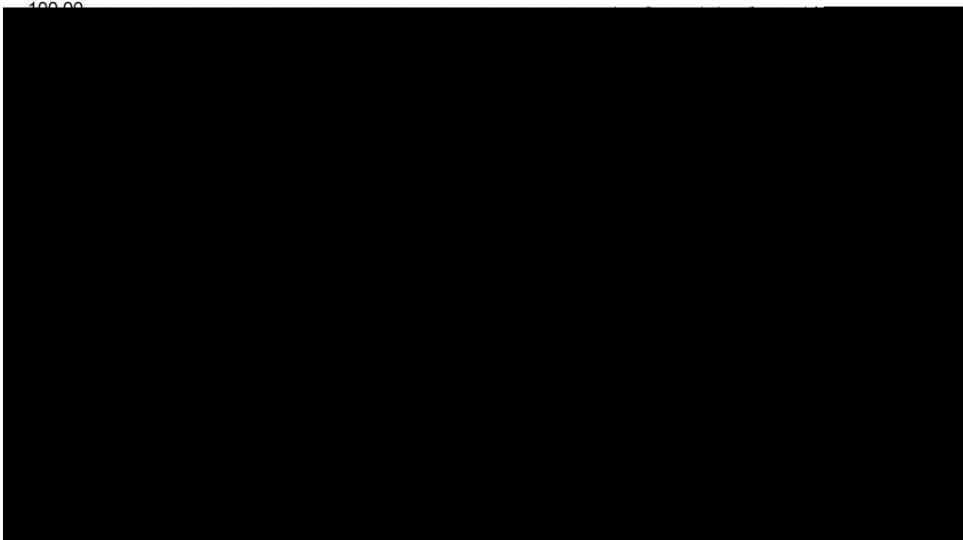


Figure 19: Data from Oxoby and Spraggon (2008).





## 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.

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:

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

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  $u(a) = (l_1, l_2 + x)$ ,  $u(b) = (l_1 + x, l_2)$ . In words, players have incomes  $l_1$  and  $l_2$  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  $D(a) = f(l_1 + x) - f(l_1)$  and the aggregate dissatisfaction of  $b$  is  $D(b) = f(l_2 + x) - f(l_2)$ . Then if  $l_1 > l_2$ , we have  $D(a) < D(b)$ . This means that consequence  $a$ , where

## 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  $\langle hN, C_1, u_1, D^1 \rangle$  and  $\langle hN, C_2, u_2, D^2 \rangle$  as separate and possessing their own norm functions  $h_{C_1}$  and  $h_{C_2}$ . In order to compare these norm functions on  $C_1 \cap C_2$ , we need to find some common ground, since  $h_{C_1}$  and  $h_{C_2}$  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  $h_{C_i}(c_i) = 1$ , 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.<sup>2</sup> 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  $\langle hN, C_1, u_1, D^1 \rangle$  and  $\langle hN, C_2, u_2, D^2 \rangle$ , call  $h_{C_i} : C_i \rightarrow [0, 1]$  defined as  $h_{C_i}(c) := \frac{\bar{D}^i(c) - m_i}{x - m_i}$  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  $h_{C_1}$  is calculated as  $\frac{\bar{D}^1}{x - m_1}$ , which is normalized from the interval that covers dissatisfactions in both environments to  $[0, 1]$ .

**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,  $h_C(\text{do nothing}) = \frac{1}{3}$  and  $h_C(c) = 1$  for the other consequences. From Definition 2 we obtain  $h_{C_1}(\text{do nothing}) = \frac{1}{3}$  and  $h_{C_1}(\text{attention}) = 1$ . Thus, the appropriateness of doing nothing is higher when there are few options to help, exactly as our intuition

## D Norm-Dependent Utility in Games with Observable Actions

Let a tuple  $G = (N, C, u, D, H)$  be an *extensive form game with observable actions*, where  $(N, C, u, D)$  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  $G$  is a standard game with utilities being the material payoffs or consumption utilities.

Let us define some notation.  $h = (a^1, a^2, \dots, a^t)$  represents a history of length  $t$  where  $a^t = (a_1^t, \dots, a_N^t)$  is a profile of actions chosen at stage  $t$ ,  $1 \leq t \leq \infty$ . Each history  $h$  becomes commonly known to all players once it occurs. Empty history  $f \in G \subseteq H$  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  $h \supseteq C \subsetneq H$ , 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

$r_i^{a_i}$  is positive only for players who chose the actions inconsistent with all consequences in  $M_{p(h)}$ . Let  $V = \{i \mid r_i^{a_i} > 0\}$  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;<sup>4</sup> and 3) the payoff  $v$

In this section we show how to incorporate norms assuming that punishment can be exercised outside the game. We start with the same game  $G$  as before and the norm function  $h_C$  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  $h_C$  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  $m_i^{a_i}$  in the same way as in the previous section only with  $h^{p(h)} = h_C$  on its domain.  $m_i^{a_i}$  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  $i$

## E Proofs

**Proof of Proposition 1.** Consider an environment  $(N, C, u, R)$  and two consequences  $c_1, c_2 \in C$  with  $u_i(c_1) > u_i(c_2)$



## 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.