A Theory of Injunctive Norms[*]

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

October 17, 2023

Abstract

Social behavior depends on norms that reflect complex interactions between self-interest, own- and others' expectations, and contextual factors. We present an axiomatic model that derives context-dependent norms from assumptions about agents' cognition, preferences, and constraints. Rooted in plausible moral psychology, the model integrates self-interested normative judgments derived from considering the entire set of feasible outcomes with empathy-driven consensus that incorporates the self-interest of all parties. The model sheds light on the sources and diverse manifestations of norm-influenced behaviors; it clarifies the prosocial nature of norms and a key source of their context-dependence, offering precise, testable explanations for norm-driven behavior.

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

[*] This is a substantially revised version of an earlier working paper of the same title that also integrates elements of a paper titled "A Theory of Moral Reasoning". A section of the earlier paper focused on the relationship between norms and punishment has been turned into a new paper (Kimbrough and Vostroknutov, 2023b). We would like to thank Valerio Capraro, Hartmut Kliemt, Sébastien Pouget, David Rojo-Arjona, Nicole Saito, Matthias Stefan, Elias Tsakas, and participants in the Works-in-Progress seminar at Chapman University's Smith Institute for Political Economy and Philosophy, the ESA-World meetings in Vancouver, the PPE Summer School at Chapman University, and seminars at GSBE-ETBC Maastricht, Ohio State, George Mason, Konstanz, Max Plank Institute-Bonn, UC-Davis, Virginia Tech, San Jose State, and Arkansas 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 behavior is often *other-regarding*. 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, behavioral economists have developed a wide variety of models meant to capture diverse motives that extend beyond mere selfishness; such models suggest that people are variously motivated by Pareto improvements, efficiency, equality, maximin, reciprocity, guilt aversion, lying aversion, anger, and so on.[1] Further experimental evidence also reveals that behavior is often *context-dependent*. For example, simple changes to the choice set are sufficient to cause people to switch from one motive to another (e.g., List, 2007; Engelmann and Strobel, 2004; Galeotti et al., 2019). However, the mapping between motive and context remains outside of the scope of most behavioral models (Vostroknutov, 2020; Kimbrough, 2022).

To account for other-regarding and context-dependent behavior in a unifying framework, economists have recently proposed that decisions reflect an intrinsic desire to adhere to (injunctive) social norms (Cappelen et al., 2007; López-Pérez, 2008; Kessler and Leider, 2012; Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2016). Injunctive norms reflect shared agreement about the social appropriateness (and inappropriateness) of various outcomes. Models of norm-dependent preferences thus 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

[1] For example, Charness and Rabin (2002); Fehr and Schmidt (1999); Dufwenberg and Kirchsteiger (2004); Engelmann and Strobel (2004); Battigalli and Dufwenberg (2007); Abeler et al. (2019); Battigalli et al. (2019b).

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, 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 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 testable predictions. We show that our model can account for a wide variety of other-regarding and context-dependent social behaviors that are otherwise difficult to explain within a single existing model.

In our framework, we use an axiomatic approach to directly model how norms emerge from a plausible moral psychology: 1) agents' normative judgments of outcomes are rooted in self-interest, and 2) the capacity for empathy allows agents to imagine and incorporate others' (similarly self-interested) normative judgments into an agreed-upon norm. We start with the assumption that each agent's fundamentally self-interested desire to achieve better outcomes for himself generates *dissatisfaction* when less preferred outcomes are achieved. Each outcome has a consumption value associated with it, and outcomes worse for me than others that might have attained evoke dissatisfaction. This kind of counterfactual reasoning generates a natural context-dependence: my dissatisfaction with one outcome depends on the set of other outcomes that might have occurred instead.

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 summed across individuals, and the least socially appropriate consequence is the one that maximizes aggregate dissatisfaction. Agents are other-regarding not directly, in the sense of getting utility from others' utility or the distribution thereof, but indirectly, insofar as empathy allows norms to account for the dissatisfaction of all.

Our modeling approach draws on a rich tradition from moral philosophy 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 conduct for others, with "fellow-feeling" allowing us to more-or-less understand how others might feel should a particular outcome 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).[2] The key to our model is the assumption that what others will go along with depends on the set of counterfactual outcomes available to them.

The key implication is that, because an individual's evaluation of each outcome depends on how it compares to *all* other outcomes, the resulting norms are *radically-context-dependent*. We show analytically that the context-dependent norms derived from the model have a number of intuitively appealing properties: 1) the most appropriate outcome is always Pareto optimal, respecting a key tenet of welfare economics; 2) as a result, the norm favors efficient equilibria in coordination games; 3) under scarcity, the most appropriate outcome always involves compromise such that no agent ought to receive the whole pie; 4) in social dilemmas, norms favor cooperative outcomes so long as non-cooperative outcomes are not substantially more efficient in terms of consumption value; 5) positive reciprocity often emerges as a norm in extensive form games; and 6) there exist "constant-norm" environments in which all outcomes are equally appropriate (e.g. tournaments, competitive markets) such that selfishness is consistent with norm-following.

Then, 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 in which behavior is known to be other-regarding and context-dependent. We show that changes in behavior across contexts track the predicted changes in norms under our model.

We consider two 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 and 2) games that differ from one another only by the addition or subtraction of possible outcomes. 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.

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 show that the theory predicts how agents select one Pareto optimum among many as normative and in particular why the equal split minimizes aggregate dissatisfaction in the standard version of the game. Analyzing the games studied by Engelmann and Strobel (2004) and Galeotti

[2] One interpretation of our theory could be in terms of a contractarian approach to morality, in which moral rules are established 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.

et al. (2019), 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 a sufficiently concave value function over money when computing individual dissatisfaction. Thus, we highlight how our model connects to the social preferences literature – helping to explain why measured social preferences vary across contexts.

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 vein, 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 through our model, as the dissatisfaction with one outcome is directly affected by the introduction (or removal) of another.

For simplicity, we have thus far assumed 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. In this sense, *the basic theory is addressed to the spare contexts typically studied in economics experiments.* However, the axioms do not preclude a situation in which different individuals' dissatisfactions are weighted differently in computing the norm. For example, kin and other in-group members might naturally count more than strangers or other out-group members in our moral calculus.

We illustrate this intuition by showing that the model can account for norms of differential treatment of in- and out-groups if, when aggregating dissatisfaction across individuals to define the norm function, we weight the dissatisfaction of others in proportion to their relative 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 draw on the experiments reported by Chen and Li (2009) in which the relative weight on the in- and out-groups can be estimated from prior data (via a within-subject experimental design) and then used to make predictions about play in subsequent games, which receive strong support.

In sum, we introduce and illustrate the interpretive power of a model of other-regarding and context-dependent injunctive norms. We construct the model from axioms that emphasize the psychological bottom-up construction of norms, combining a notion of normative dissatisfaction with a capacity for empathy. Crucially, individual dissatisfaction is derived from self-interest such that norms vary only with agents' preferences over consumption bundles and the choice set in which those are embedded. In particular, we assume that agents are dissatisfied with a particular outcome to the extent that there exist other outcomes that would yield a higher counterfactual consumption value. Empathy, then, serves to temper this self-interest by allowing agents to recognize that others feel similar dissatisfaction, and the resulting norm ranks outcomes according to the aggregated dissatisfaction of all interested parties at each outcome.

To our knowledge, ours is the first model attempting to account for the *content* of injunctive norms in laboratory environments endogenously, as a function of the set of available outcomes. 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, to work out the rich implications of the model for norms in a number of widely studied games, and then to take the model to data.

Existing theories in which normative considerations shape choices differ from our approach in that they fail to provide an account of how injunctive norms emerge from the cognition, preferences, and constraints of individuals. In models of social preferences (e.g., inequality aversion, efficiency-seeking, maximin) and in models of image concerns (e.g. Bénabou and Tirole, 2011) and psychological game theory (e.g. Battigalli et al., 2019a), the injunction comes from the modeler. Social preference theories bake a particular injunctive norm directly into agents' preferences; image models and psychological games more subtly depend on normative judgments made by the modeler regarding which actions harm/help one's image or evoke a particular psychological response.

Endogenizing the norm to the choice context allows one to make testable predictions about which injunctive norm will operate in which context and which actions will be the targets of resentment (and hence punishment; see Kimbrough and Vostroknutov, 2023b), and thus we hope our model complements work on image and psychological game theory by providing grounds for the normative assumptions made therein. In a followup paper, we offer a framework for defining dissatisfaction functions and the resulting norms more generally, and we discuss the extent to which the context-dependence of behavior that has been observed in lab experiments can be captured in a model that exogenizes the norm, e.g. by reference to some ideal (Kimbrough and Vostroknutov, 2023a).

In the next section, we introduce our model and describe in detail its implications for norms across a variety of economically-relevant contexts. In section 3, we take the model to existing experimental evidence to illustrate its interpretive power. Finally, in section 4 we summarize, discuss some limitations of the model, and consider directions for future research.

2 Model

We provide a set of axioms rooted in moral psychology that endogenize injunctive norms to the choice context. We start from the premise that individual normative judgments are ultimately founded on individual interests and that shared norms tend to arise which, through empathy, balance the consideration of these interests across agents in each choice set.[3] In particular, we as-

[3] To illustrate the premise, every parent knows that children learn to understand the concept "mine" before they learn to understand "yours". Children naturally understand that it is in their interest to have others respect their

sume that all agents are dissatisfied when they are unable to achieve outcomes that are selfishly better for them. Since empathy enables each agent to imagine how others are also motivated by their own interests, the resulting norm asks each agent to temper his own self-interest, bringing it down to a level that others "can go along with" (Smith, 1759; Smith and Wilson, 2019).

Our model begins with the assumption that individual normative judgments are fundamentally comparative. We define individual normative judgments via "dissatisfaction" functions that evaluate how dissatisfied agents are with a particular outcome *because of* how it compares to other feasible outcomes. Our axioms define how the dissatisfaction function is constructed and how individual dissatisfaction functions are aggregated into a shared injunctive norm.

We assume that dissatisfactions, and the resulting injunctive norms, depend on the final outcomes of a game and not on its strategic structure defined by a sequence of moves, information sets, etc. Thus, we start with a large set $\mathcal{C}$ of all possible consequences, N players, and a consumption value function $u : \mathcal{C} \to \mathbb{R}^N$, which for each consequence defines a vector of consumption values (payoffs). Suppose that the image of u is $\mathbb{R}^N$ so that all payoffs are possible. We will work with *finite* subsets of $\mathcal{C}$ (called contexts), with a typical context denoted by $C \subset \mathcal{C}$. We think of C and the collection of associated payoff vectors $u[C]$ as the set of all feasible allocations in the context of a choice problem or game.

We first introduce axioms on functions $d_i : \mathbb{R}^2 \to \mathbb{R}_+$ that define *dissatisfactions* for each player $i \in N$. Namely, $d_i(u_i(x), u_i(y))$ stands for the dissatisfaction that player i feels when x is attained and y was available. Then we introduce axioms on functions $D_i(x \mid C)$ that capture player i's *total dissatisfaction* from consequence $x \in C$ because of all other consequences in C. Together, these provide us with a theory of individuals' normative evaluations of each consequence as a function of other consequences in the choice set. In the third step, we define an *aggregate dissatisfaction function* $D(x \mid C)$ that creates a composite of the dissatisfaction of all interested players. This function yields a normative comparison of all feasible consequences which can be directly translated into a "normative valence" of each element x in C, for use in a norm-dependent utility function.

2.1 Axioms on Dissatisfaction Functions d_i

Our axioms on d_i capture the intuitive notion of comparative dissatisfaction, in which the normative evaluation of a consequence depends on how it compares to another consequence in terms of consumption value. A1 and A2 guarantee that dissatisfaction is measured in the same units as consumption value. A3 restricts the focus to negative comparisons (*dis*-satisfaction), and A4 postulates that dissatisfactions are comparable across people. A1 and A2 capture the idea that

claims to property, but they begrudgingly learn that others have similar interests. The convention of property is then founded upon mutual recognition of these interests.

people's normative evaluations of a consequence are motivated by self-interest, and A4 relies on empathy to introduce the possibility of interpersonal comparison.

A1 $\forall i \in N, \forall x, y \in \mathcal{C}, \forall a \in \mathbb{R}, \;\; d_i(u_i(x) + a, u_i(y) + a) = d_i(u_i(x), u_i(y)).$

A1 states that adding a constant to the consumption values being compared does not change the dissatisfaction with one outcome because of the possibility of another. So, if player i gains or loses the same amount of consumption value in all consequences, then her dissatisfaction is unaffected. A1 ensures that d_i can be expressed as a function of the difference of consumption values.

A2 $\forall i \in N, \forall x, y \in \mathcal{C}, \forall \alpha \in \mathbb{R}$ with $\alpha > 0 \;\; d_i(\alpha u_i(x), \alpha u_i(y)) = \alpha d_i(u_i(x), u_i(y)).$

A2 states that if all consumption values or payoffs are multiplied by a positive constant, then the dissatisfactions are also multiplied by the same constant. This ensures that dissatisfactions are proportional to consumption values in a linear way, thus connecting the two concepts. We could have assumed some non-linear, say concave, relationship between dissatisfaction and consumption value. However, we already allow consumption value to be a non-linear function of payoffs. A2 reflects an idea that all non-constant marginal effects of payoffs are already encoded in functions u_i.

A3 $\forall i \in N \; \forall x, y \in \mathcal{C}$ with $u_i(x) \geq u_i(y)$ we have $d_i(u_i(x), u_i(y)) = 0.$

A3 introduces an asymmetry between prospective gains and prospective losses. In particular, A3 says that players do not feel dissatisfaction with a superior outcome because of inferior consequences. In other words, any positive sentiment that a player may feel because there exist some inferior consequences (as in, "hey, it could be worse") does not influence the dissatisfaction with the superior consequence.[4] It is important to note that we are not assuming that players cannot feel such sentiment in these circumstances, only that this sentiment does not influence the computation of injunctive norms. Finally, we assume non-triviality and equivalence of dissatisfactions across players:

A4 $\forall i \in N \;\; d_i(0, 1) = 1.$

The following proposition establishes the functional form of d_i equivalent to axioms A1-A4.

Proposition 1. The following two statements are equivalent:

1. d_i satisfies A1-A4;
2. $d_i(u_i(x), u_i(y)) = \max\{u_i(y) - u_i(x), 0\}.$

[4] This idea goes back at least as far as Smith (1759) and has a strong empirical foundation (Tversky and Kahneman, 1981). For simplicity, we assume that the asymmetry is severe, but the implications of our model still go through if we assign non-zero but smaller weight to the "gratitude" that arises from avoiding inferior consequences.

Proof. See Appendix C.

This function computes the *dissatisfaction* that player i feels about the consumption value of some consequence x *because of* the possibility of consequence y. 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 $d_i(u_i(x), u_i(y))$ because y could have attained instead. Dissatisfaction is positive when y brings player i more consumption value than x and zero otherwise. The function d_i only defines how i normatively compares two consequences for himself to one another. It does not compare across people within one consequence, as is typical in the social preferences literature. Norms become social in the process of aggregation, which we will discuss below in Section 2.3.

2.2 Axioms on Total Dissatisfaction Functions D_i

Our axioms on D_i define the total dissatisfaction of player i associated with a single consequence x. A5 postulates that individuals feel dissatisfaction only when other consequences are feasible; in a singleton choice set, there is nothing to be dissatisfied about. A6 says that all counterfactual outcomes are treated equally regardless of which context they appear in. This axiom is particularly important because it ensures that the total dissatisfaction of any consequence x can only depend on the consumption values of the other outcomes in the context, and it ensures that when a new consequence y is added to a context, its effect on the total dissatisfaction with x only depends on the value difference between y and x. Adding y to a context may also impact the total dissatisfaction with another consequence z, but this change in the total dissatisfaction with other consequences has no bearing on the change in the dissatisfaction with x. We think of A6 as capturing another basic sentiment, namely that player i feels dissatisfaction whenever a new possibility y appears, that could give i a higher payoff.

A5 $\forall i \in N\ \forall x \in \mathcal{C} \quad D_i(x \,|\, \{x\}) = 0.$

Axiom A5 states that if only one consequence is in the choice set, then there is nothing to be dissatisfied about, so the dissatisfaction of each player i is zero. Although A5 may sound trivial, it rules out any situations in which players are dissatisfied due to specific properties of an allocation x given the choice set $\{x\}$.[5]

A6 $\forall i \in N\ \forall C \subset \mathcal{C}, \forall x \in C, \forall y \in \mathcal{C} \backslash C$

$$D_i(x \,|\, C \cup \{y\}) = D_i(x \,|\, C) + d_i(u_i(x), u_i(y)).$$

A6 defines the total dissatisfaction function of player i at $x \in C$. It says that given any set of consequences C, any consequence x in this set, and any consequence y outside C, the dissatisfaction with x in the augmented set $C \cup \{y\}$ equals that of x when y is not in the set plus some non-negative number $d_i(u_i(x), u_i(y))$ that depends *only* on i's payoffs in x and the payoffs in the added consequence y. Moreover, the higher the payoffs in y the higher is the dissatisfaction (guaranteed by the assumptions on d_i above). The important implications of this definition are 1) that i's dissatisfaction with x in different sets of consequences is connected and 2) that the amount by which i's dissatisfaction changes when a consequence y is added only depends on the payoffs in x and y and does not depend on the characteristics of C.

The following result connects the axioms above and the representation that we test below.

Proposition 2. The following two statements are equivalent:

1. D_i satisfies A5-A6*;*
2. D_i can be expressed as $D_i(x \,|\, C) = \sum_{y \in C \backslash \{x\}} d_i(u_i(x), u_i(y))$.

Proof. See Appendix C.

The proof operates by constructing the total dissatisfaction of a consequence x from the singleton set x and incrementally adding the other consequences in C. The axioms guarantee that the total dissatisfaction with $x \in C$ is the same regardless of the order in which consequences are added, and this explains why the total dissatisfaction with x is the sum of the dissatisfaction with all other consequences in C. This means that a low value consequence results in more (less) dissatisfaction the larger (smaller) is the set of counterfactual consequences that yield a higher value. 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.

To see why including the dissatisfaction generated by *all* other consequences is important, consider an alternative model in which dissatisfaction is defined only in terms of the highest-

[5] This makes our model incompatible with social preference utility specifications where the utility of a player may depend directly on the payoffs received by other players at the same consequence. A form of social preferences, in the common meaning of the term, could be introduced if in A5 we assumed that dissatisfaction is not zero, but rather depends on x. However, we deliberately eschew this path, as we believe it is more economical to understand social preferences as an epiphenomenon of norm-dependent preferences (a view we also spell out in Kimbrough and Vostroknutov, 2016; Kimbrough, 2022). Indeed, one of our goals is to show how particular kinds of social preferences (and predictable variation in social preferences across contexts) can be explained via the model presented here.

valued alternative. Mathematically, this could be expressed as $D_i(x \mid C) = \max_{y \in C} d_i(u_i(x), u_i(y))$.[6] Applying the aggregation procedure described in the next subsection, it is straightforward to show that this alternative definition of total dissatisfaction results in substantially less context-dependence than our formulation. For example, it always ranks consequences according to efficiency and does not differentiate among consequences with a fixed sum of payoffs. In Kimbrough and Vostroknutov (2023a) we analyze alternative methods of defining dissatisfaction functions in more detail, comparing our model of injunctive norms to a variety of less radically context-dependent "moral rules."

2.3 Axioms on Aggregate Dissatisfaction Function D

Next, we aggregate the dissatisfactions across players and define D, aggregate dissatisfaction. This aggregation procedure combines individual normative judgments (D_i) to generate a shared normative agreement that assigns an appropriateness to each consequence. We start by assuming that $D(x \mid C)$ is a function of $D_i(x \mid C)$ for all $i \in N$. Specifically, we assume that $D(x \mid C) = G(D_1(x \mid C), ..., D_N(x \mid C))$, where $G : \mathbb{R}^N \to \mathbb{R}_+$ is increasing in all arguments.

A7 $G(0, ..., 0) = 0$.

A7 simply states that if each player feels the lowest dissatisfaction (i.e. zero), then the aggregate dissatisfaction is also minimized and equals zero.

The last axiom (A8) defines how changing the dissatisfaction of one player changes aggregate dissatisfaction. For generality, and to allow us to model interactions between individuals who differ in the priority assigned to them in normative judgments, we assume that players have social weights $(\omega_i)_{i \in N}$, where $\omega_i \in \mathbb{R}$. These weights determine how much each player's dissatisfaction counts in the computation of aggregate dissatisfaction, and they can represent power, social status, in/outgroup relationships, kinship, or their combination.[7]

A8 $\forall i \in N \; \forall D_1, ..., D_N \in \mathbb{R}_+ \; \forall a_i \geq -D_i \quad G(D_i + a_i; D_{-i}) = G(D_i; D_{-i}) + \omega_i a_i$.

The proposition below puts all the axioms together.

Proposition 3. The following two statements are equivalent:

1. d_i satisfies A1-A4, D_i satisfies A5-A6, D satisfies A7-A8.
2. D can be expressed as

$$D(x \mid C) = \sum_{i=1}^{N} \omega_i D_i(x \mid C) = \sum_{i=1}^{N} \sum_{y \in C} \omega_i \max\{u_i(y) - u_i(x), 0\}$$

[6] This is similar to the formulation proposed in Cox et al. (2018).

[7] We find this approach appealing as real-world normative disagreements can often be fruitfully understood in terms of disagreements about these weights.

Proof. See Appendix C. [8]

The function D captures the dissatisfaction of all players for each possible consequence in a game. The aggregation of dissatisfaction across individuals reflects our assumption that aggregate dissatisfaction of the players depends only on their individual dissatisfactions. This aggregation is intended to reflect the capacity for empathy, with individuals applying their knowledge of how others would feel at a given consequence to agree upon a normative ranking. We then define the normatively best outcome as the one that minimizes aggregated dissatisfaction.

2.4 The Norm Function

We assume that the *normative valence* of a consequence x is inversely proportional to its aggregate dissatisfaction. 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, 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 $\eta : C \rightarrow [-1, 1]$, defined as*

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

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

In this definition, η is simply the negative of aggregate dissatisfaction, normalized to the interval $[-1, 1]$. 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, but normalization is justified since $D(x \mid C)$ and $\alpha D(x \mid C)$ for any $\alpha > 0$ are equivalent representations. Moreover, normalization becomes important when different norm functions are compared (as in Appendix B, for example). Thus, we assume that the consequence x with $\eta(x) = 1$ is the most socially appropriate (the norm) and the one with $\eta(x) = -1$, the least socially appropriate. If all consequences

[8] In the case with only one player ($N = 1$) and keeping in mind the result of Proposition 1, the normative ranking of outcomes implied by our axioms is consistent with *regret avoidance* similar in flavor to that considered by Fioretti et al. (2022). In the canonical treatment of regret (e.g., Loomes and Sugden, 1982), it is defined as a negative feeling associated with the *unchosen* alternative. The form of "regret" considered here and in Fioretti et al. (2022) can be felt about any counterfactuals, regardless of their current attainability with choice. However, our model does not reduce choice to the maximization of a regret averse utility function; rather, the aggregated injunctive norms generated by our model can be thought of as preferences of a regret-averse player who cares also about the regrets of other players, and norm-following is traded off against maximization of consumption value in choice.

have the same aggregate dissatisfaction, then we assume that $\eta(x) \equiv 1$ 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 B). We next illustrate some properties of η.

2.5 Some Properties of the Norm Function

In this section we provide some theoretical results on the properties of the norms derived from our model. This serves to establish some regularities, highlighting that the model has structure despite its radical context-dependence, and provides intuition for the kinds of context dependence that emerge naturally from it.

2.5.1 Norms and Consumption Value

The first thing to note about the norm function is that its evaluation of consequences will also depend on the function defining agents' consumption values, since that is the criterion by which they evaluate their dissatisfaction. For example, if the consumption value of money is linearly increasing, then dissatisfactions of all players will not depend on their wealth, but if it is logarithmic in money, then dissatisfaction of poorer individuals will be larger than that of richer individuals. Thus, norms will be skewed towards favoring poorer individuals in the logarithmic case. If the value function is sufficiently concave, then norms will approximate maximin (Rawls, 1971).

2.5.2 Norms and Pareto Optimality

Secondly, injunctive norms derived from the model have an important property: in *any* set of consequences, a Pareto dominated consequence is always normatively inferior to another that Pareto dominates it (i.e., it always generates more aggregate dissatisfaction). Thus, injunctive norms derived from the model always satisfy the Pareto optimality criterion. We formulate this as a proposition.

Proposition 4. *In any context $C \subset \mathcal{C}$, if $x \in C$ Pareto dominates $y \in C$ then $D(x|C) < D(y|C)$.*

Proof. See Appendix C.

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. Thus, although all dissatisfaction-minimizing consequences are Pareto optimal, the converse is not true. In fact, we prove the following relatively general result that shows that the normatively best consequences under our model are never such that one player gets the maximum payoff; that is, everyone is expected to compromise to some degree. This holds under the condition that there is only one consequence, different for each player, where their maximum payoff is reached. This *Scarcity Condition*, as we call it, is satisfied in very general classes of social dilemmas (for example, various asymmetric versions of Dictator, Trust, and Public Goods games) and can be seen as a kind of resource constraint in a strategic environment. We provide formal definitions and proofs of this result in Appendix D. Here we simply offer a sketch of the proposition to build intuition.[9]

Proposition 5. (Compromise Theorem) *Suppose that C is an N-dimensional convex polytope in $\mathbb{R}^N$ that satisfies the Scarcity Condition. Then, the normatively best consequence according to the model is such that no one player gets the maximal possible payoff.*

Proof See Appendix D.

Proposition 5 essentially says that under a certain scarcity restriction satisfied in most social dilemmas with continuous action spaces (e.g., Trust game, Public Goods games with various asymmetries), the normatively best consequence never gives maximal payoff to any one player. In other words, in such environments the model generally predicts that all players need to compromise and sacrifice something for the sake of achieving a normatively best outcome. This makes intuitive sense, as the Pareto-optimal allocations that give all the resources to a single player have been a source of criticism of the Pareto optimality criterion.

Finally, we prove a different version of the Compromise Theorem for contexts involving only two players and constant efficiency, but which nonetheless provides a more specific characterization of the most appropriate allocation.

Proposition 6. *Suppose there are two players and K consequences $C = \{x_1, x_2, ..., x_K\}$ with consumption value functions $u_1 \leq u_2 \leq ... \leq u_K$ 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 $j = 1..K-1$, $D(x_{j+1}|C) - D(x_j|C) = (2j-K)(u_{j+1} - u_j)$. Thus, the midpoint consequences $x_{\frac{K}{2}}$ and $x_{\frac{K}{2}+1}$, if K is even, and $x_{\frac{K}{2}+\frac{1}{2}}$, if K is odd, have the smallest dissatisfaction.*

Proof. See Appendix C.

Proposition 6 implies that the most appropriate consequence, in cases with constant payoff efficiency over all possible allocations, is not the one that is the closest to an equal distribution of consumption value, as most models of social preferences would suggest, but rather the one

[9] While the axioms are defined over finite sets of consequences, we generalize the results on utility representation to continuous sets of consequences. We do not provide axioms for the continuous case and leave it for future research.

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 that yield very unequal values across agents, can still be considered normatively appropriate in specific contexts where most consequences give a large portion of the pie to one player. In the following example, we showcase how this works in the continuous action space of a dictator game.

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

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

This is an upward sloping parabola which is minimized at $x^* = \frac{1}{2}$. Thus, the norm function $\eta(x|C)$ is a downward sloping parabola with the equal split being the most socially appropriate consequence and the consequences $x = 0$ and $x = 1$ 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, but the equal division emerges as the most appropriate consequence because of the game’s symmetry.

To see how dissatisfaction affects the norm when the consumption values of players are asymmetric, let us assume that the receiver has a different “need” for the pie than the dictator. Suppose the receiver’s consumption value is $u_r(x) = \gamma x$, 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 $x^* = \gamma/(1 + \gamma)$, 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. □

Next, we establish some theoretical results about other commonly studied games in which normative motivations arguably play an important role in decision-making.

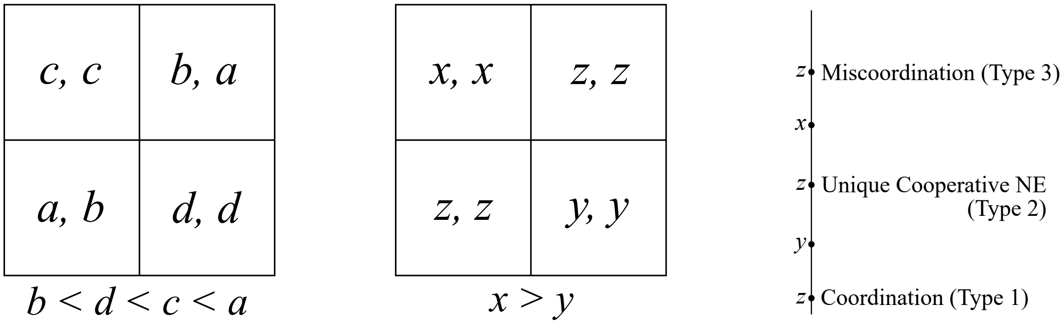
2.5.3 Norms in Coordination Games

In coordination games, the normatively best outcome depends on what other assumptions we make about payoffs. In coordination games with multiple Pareto-ranked equilibria (minimum effort games, stag hunts), the model will always select the Pareto-optimal equilibrium as the normatively best outcome. By contrast, in games with symmetric, non-Pareto-ranked equilibria (e.g., matching pennies, battle of the sexes), the model provides little guidance if players are treated identically in aggregating dissatisfaction. However, in these kinds of settings, heterogeneity across players can help resolve normative ambiguity, as such games are more likely to have a unique normatively best outcome if the players' dissatisfactions are not weighted identically with social weights ω_i in computing the norm (e.g., if it is one of the two players' birthday in the battle of the sexes).

2.5.4 Norms in Social Dilemmas

An intuitive norm for social dilemmas is that players ought to cooperate, and indeed this is frequently, though not always, consistent with the norm derived from our model. For intuition, consider a two-player Prisoner's Dilemma game played by coequal strangers. It is easy to see that the injunctive norm will generally favor the outcome cooperate/cooperate since the resulting payoffs are typically efficient and relatively egalitarian in most Prisoner's Dilemma games studied by economists. The exception arises if one player's payoff from unilateral defection is sufficiently high that the injunctive norm favors the outcome cooperate/defect (or defect/cooperate); this is because efficiency considerations can dominate in such extreme cases.

Example 2. Prisoner's Dilemma. Consider the Prisoner's Dilemma with material payoffs a, b, c, d as shown on the left graph of Figure 1. 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

$x > y$

Figure 1: 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 1. 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 1). 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. □

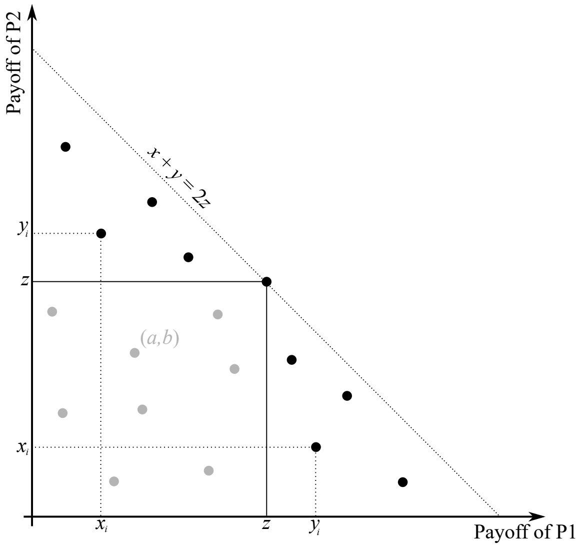


Figure 2: Symmetric two-player social dilemma.

Next, we provide formal analysis that generalizes this point, considering a class of symmetric two-player social dilemmas in which the norm corresponds to the most efficient outcome—even though for both players, there exist other consequences that bring them higher consumption value. Consider the set of payoff allocations for two players shown in Figure 2. Here we identify consequences with points on $\mathbb{R}^2$. Point (z, z) represents the most efficient symmetric allocation, or the cooperative outcome. Allocations (x_i, y_i) and (y_i, x_i) that are weakly less efficient ($x_i + y_i \leq 2z$) and satisfy $x_i \leq z$, $y_i \geq z$ for all i represent the possible unilateral defection outcomes that give more payoff to the defector than in the cooperative outcome and less to the other player. Finally, arbitrary points (a_i, b_i) with $a_i \leq z$ and $b_i \leq z$ for all i represent the possible mutual defection outcomes.

This is a rather general class of games that includes most Prisoner's Dilemmas (with efficiency of the cooperative outcome restricted to be at least as high as the defect-cooperate outcome), the two-player Public Goods game, and even the Dictator game as a special case. The following proposition shows that the allocation (z, z) is maximally appropriate.

Proposition 7. *For 2 players consider the set of payoff vectors that consists of 1) point (z, z); 2) n pairs of points (x_i, y_i) and (y_i, x_i) with $x_i + y_i \leq 2z$ and such that $x_i \leq z$ for all $i = 1..n$ and $z \leq y_1 \leq y_2 \leq$*

... $\leq y_n$; 3) any finite number of other points (a_i, b_i) with $a_i \leq z$ and $b_i \leq z$. Then (z, z) has the smallest dissatisfaction.

Proof. See Appendix C.

To provide additional intuition about the kind of behavior implied by the injunctive norm in social dilemma games, consider Figure 3 that illustrates the aggregate dissatisfaction functions in a 2-player Public Goods game (using parameters derived from Fehr and Gächter, 2000) and a Trust game (using parameters from Berg et al., 1995). On both graphs the points on the 2D plane are the payoffs that players can obtain and the color encodes the aggregate dissatisfaction, with dark red being the outcome with the least dissatisfaction and dark blue the outcome with the most dissatisfaction.

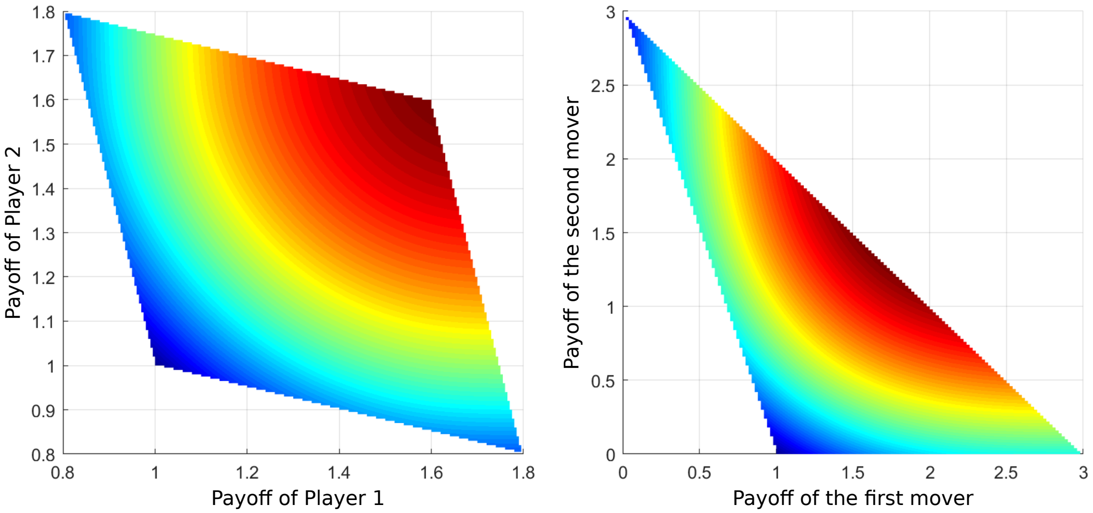


Figure 3: **Left:** The norm function in a 2-player Public Goods Game. **Right:** The norm function in a Trust Game. Dark red shows the most appropriate consequence and deep blue shows the least appropriate one.

The most appropriate consequence in the Public Goods game is for both players to contribute the whole endowment, which follows from Proposition 7, and the most inappropriate consequence is to contribute nothing. The normatively best outcome maximizes both efficiency and equality in this case.[10]

In the Trust game the most appropriate consequence is for the first mover to send everything to the second mover (1 token), and for the second mover to return *slightly more than half* of the resulting amount (second mover returns 1.66 tokens and keeps 1.34 tokens). This should not come as a surprise since the payoffs in the Trust game are not symmetric and do not satisfy the assumptions of Proposition 7. It is also worth noting that for any choice of the first mover, the

[10] Technically, Proposition 7 holds for finite sets of allocations and does not apply to the allocations sets in Figure 3 that have the power of the continuum. However, given the continuity properties of approximations of Lebesgue integrals, it is relatively clear that a limit version of Proposition 7 for continua can be formulated. We leave this for future research.

norm (the allocation with the smallest dissatisfaction) prescribes that the second mover ought to return around half the resulting money (which is equal to the amount sent times 3) back to the first mover. That is, the norm favors what looks like positive reciprocity.

2.5.5 Positive Reciprocity In Extensive-Form Games

In this section we demonstrate how positive reciprocity can emerge in extensive-form games. By positive reciprocity we mean the tendency to return a favor. We do not consider negative reciprocity here; we model punishment in a followup paper (Kimbrough and Vostroknutov, 2023b).

Consider a 2-player game with one move, akin to the Dictator game, but without restrictions on payoffs. Suppose the resulting payoff vectors are given by (p_{1i}, p_{2i}) where i ranges over some set E. Now, suppose that after Player 1 makes a move, the game is repeated once more but with players' roles reversed. The resulting extensive-form game represents the "gift exchange" between the two players.

The set of payoff vectors of this game can be described as $S = \{(p_{1i} + p_{2j}, p_{2i} + p_{1j}) \,|\, i, j \in E\}$ when Player 1 chooses allocation i and Player 2 chooses allocation j. This set has several important properties. First, when both players choose the same action, their resulting payoffs are equal and lie on the 45° line from the origin. Second, for each payoff vector $(x, y) \in S$ that is not on the 45° line, there is a symmetric vector $(y, x) \in S$. Third, take action a^* of either player which leads to the intermediate payoffs (p_{1i}, p_{2i}) with the highest efficiency $(p_1^*, p_2^*) = \text{argmax}_{(p_{1i}, p_{2i})}\, p_{1i} + p_{2i}$. In the following proposition we show that choosing a^* by both players leads to the outcome which is the norm, or to the outcome with the smallest dissatisfaction.

Proposition 8. *In the gift exchange game the choice of action a^* by both players, which leads to payoffs $(p_1^* + p_2^*, p_1^* + p_2^*)$, has the smallest dissatisfaction.*

Proof. See Appendix C.

Proposition 8, which makes use of Proposition 7, has some illuminating implications. In any environment where players can choose allocations for themselves and another player and where the roles are often reversed, for example food sharing in small groups, we should expect that norms create a pattern of reciprocation with the same action a^*. This pattern is most pronounced when players choose in a repeated Dictator game, which can be seen as a choice of how much resources (food) to give to another player. Notice that here any action of Player 1 or 2 can be counted as action a^* since all intermediate allocations (p_{1i}, p_{2i}) have the same payoff efficiency. In this case, the norm (after two Dictator games) is to divide money equally between the players. This is achieved when both players choose the *same* action, no matter what this action is. In other words, for any X the most appropriate outcome is reached whenever Player 1 gives Player 2 X% of the pie and then Player 2 gives back X% when it is their turn to share. This can be described by the simple and well-known Golden Rule: do unto others as you would have them do unto you. In this sense, positive reciprocity can be seen as a regularity generated by our model of norms.

2.5.6 Constant-Norm Environments

While the model applies to a broad array of choice settings, there are some settings in which it provides no guidance about what one ought to do. That is, there exist environments in which the normative evaluation implied by the model is constant across all feasible consequences. Consider an environment $\langle N, C, u, D\rangle$. Let us call it *constant-norm* if $D(x|C) \equiv d$, where d is some constant. We do not (yet) have a characterization of the set of constant-norm environments in terms of payoffs, and in fact, we suspect that there are no intuitively interpretable constraints that define them. For example, the environment with two players defined by $C = \{a, b, c\}$ and $u(a) = (0, 3)$, $u(b) = (3, 0)$, $u(c) = (1, 1)$ is constant-norm, but any infinitesimal change in any payoff will make aggregate dissatisfactions of some consequences different and thus not constant-norm. Nevertheless, there is an important class of constant-norm environments that we would like to describe. These are the environments that have a structure reminiscent of a tournament.

Let us call an environment $\langle N, C, u, D\rangle$ a *tournament* if there are prizes $x_i \in \mathbb{R}$ for $i = 1..N$ such that C is the set of all 1-to-1 functions $\zeta : N \to \{x_1, x_2, ..., x_N\}$ and $u(\zeta) = (\zeta(i))_{i \in N}$. In words, in a tournament N players "compete" for prizes in the set $\{x_1, x_2, ..., x_N\}$. Each prize is assigned to some player, and the set of consequences consists of *all* such assignments, as in professional sports or poker tournaments. Note that the prize could be winner-take-all such that an indivisible object will be assigned to one of the players; it can be a few prizes of different amounts; or anything else. An important property of tournaments is that they are constant-norm. We prove this result in a proposition.

Proposition 9. *Any tournament is constant-norm.*

Proof. See Appendix C.

Thus, built into our model, which accounts for the dissatisfaction of all interested parties, is the existence of situations in which norms do not single out any consequence as more appropriate than others.

In games with normatively rankable outcomes, a model of norm-dependent utility predicts that individuals' decisions trade off consumption value maximization and adherence to norms. In constant-norm settings, the norm has no influence on choices, and individuals are expected to behave as self-interested consumption value maximizers, which seems like a reasonable prediction of players' motivation in tournaments.[11]

3 Evidence

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. Finally, we show how norms change when individuals are weighted heterogeneously in the aggregation process, by examining differences in the treatment of in- and out-group members in a setting where the weights can be estimated from data.

In what follows, we assume a utility function that takes norms as an input. Up to this point our model was purely normative, in the sense that it only described how appropriate or inappropriate the consequences of actions can be. To predict behavior, we also need to consider how appropriateness influences choices. 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 $x \in C$ as

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

where $u_i(x)$ is the consumption value of consequence x, $\eta(x \mid C)$ is the normative valence of x, and $\phi_i \geq 0$ 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 consumption value maximizer, as $\phi \to \infty$ we have player i who only cares about following norms.[12]

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,

[11] Notice as well that even though norm-following players behave selfishly in tournaments, it does not mean that in such environments they "do not care about norms." In fact, the constant norm function in tournaments implies that players who *lose* (e.g., get the smallest prize) do not find it normatively wrong, because for them all outcomes of the tournament are equally appropriate. This means that norm-following losers of a tournament do not feel disappointed with the outcome and are thus less likely to contest the results, which may be considered as an important benefit of following norms in competitive environments.

[12] 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.

keeping everything else constant, and which generate comparative static predictions that we can assess with the data.

3.1 Context-Dependent Social Preferences

We first 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: one based on valuing monetary payoffs linearly and another that values money logarithmically. 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 the linear and log cases. Figure 4 presents the norm functions for the five tasks.

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 for both linear and log. 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.

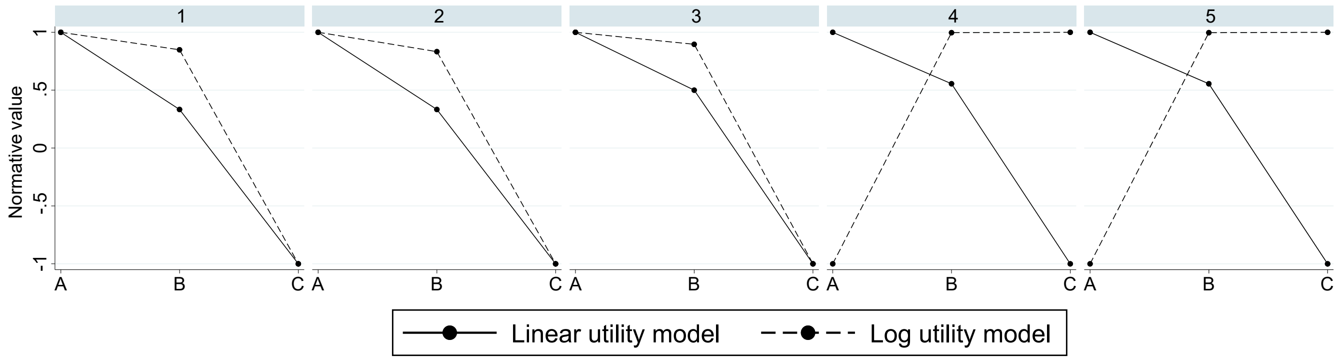


Figure 4: Normative valences in linear and log consumption value models.

In tasks 4 and 5 the situation is very different; now the payoffs of Person 3 grow in the 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 norm based on the linear value function still prescribes the choice of the most efficient allocation, the norm under the log value function 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 value functions is "correct." Rather, we see them as two ways of thinking about appropriateness in 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 value in the log 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 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. (2019) which analyzes the efficiency-equality trade-off in bargaining situations.

Case 2. Galeotti et al. (2019). 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. 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 the highest 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 an 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 5 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.

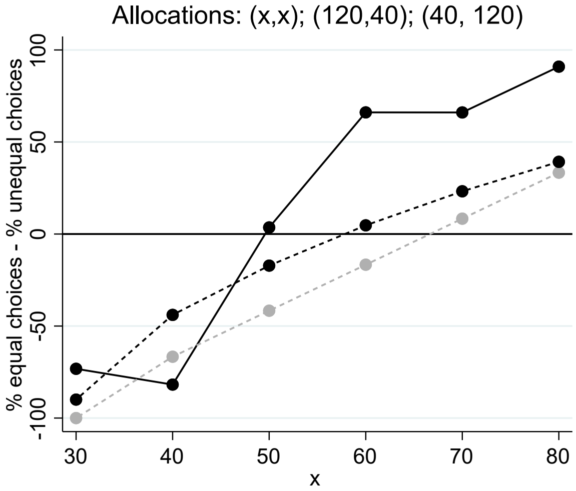


Figure 5: Solid line represents the data from Galeotti et al. (2019). Dashed lines show the predictions implied by norm functions computed under both the linear and the log model of consumption value (grey and black lines respectively).

To compare the predictions of our model with these data, we again compute two variants of the injunctive norm for each allocation: with linear and log value functions over payoffs. The dashed black line on Figure 5 shows the differences in dissatisfactions between the equal and unequal allocations under the log value function (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 value function model. In this case, linear value performs worse than log in accounting for behavior. □

This case shows that our model can capture the efficiency-equality trade-off studied in Galeotti et al. (2019). 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 5, 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. Moreover, in philosophical debates, maximin preferences (Rawls, 1971) are usually counterposed to 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 instead be derived from the single idea that normative appropriateness comes from dissatisfaction, which varies with context.

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 behavior.

We start with the give and take DGs analyzed by List (2007) (see also the experiments by Bardsley (2008) 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 changing 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).

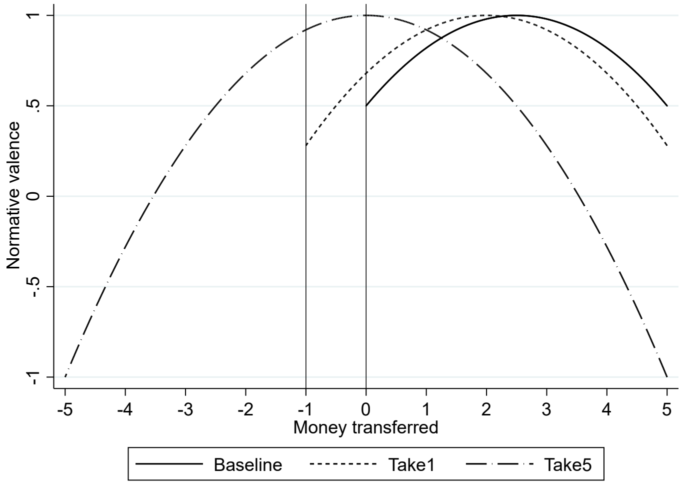


Figure 6: Relative norms in the three treatments of List (2007).

The graph on Figure 6 shows the *relative* norm functions in the Baseline, Take1, and Take5 treatments.[13] In all three cases, the aggregate dissatisfactions are calculated in the same way as for the standard DG (see Example 1). Recall that 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 Take5. This is what List (2007) reports (see Figure 10 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), 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.

Next, we analyze experimental findings of McCabe et al. (2003). 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 4. 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 7).

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

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

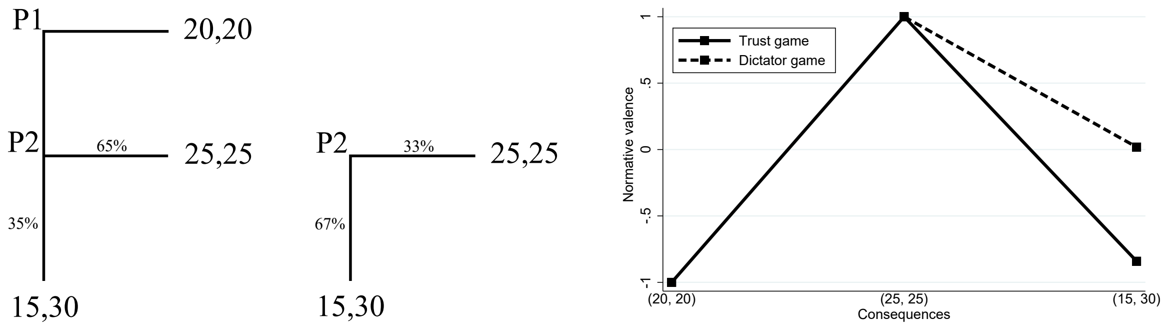


Figure 7: **Left:** the Trust and Dictator games analyzed in McCabe et al. (2003). **Right:** the normative valences of the consequences in the two games (log value function, relative norms). For the linear value function norms, see Figure 12 in Appendix A.2.

the right of Figure 7 shows the norm functions calculated with logarithmic consumption value functions and renormalized relative to each other (see Appendix B). The results are qualitatively the same with linear value functions (see Figure 12 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 Norms with Heterogeneous Weights

Finally, we show how our model can be generalized to account for heterogeneous weighting of individuals in the calculation of the norm, which can account for the 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 apply them as weights during aggregation in the model. If all the weights are equal, the model neatly captures the situation in typical lab experiments, in which all subjects belong to the same group of people (students) and have indistinguishable social status (because of anonymity). We show what happens in the model when we break this symmetry.

Suppose the set of players N is partitioned into two groups $N = \{N_1, N_2\}$. As in A8, we assume that the total dissatisfactions D_i of agents are assigned *social weights* in the aggregation. For example, the dissatisfactions of the out-group players can be discounted with a weight $\omega \in [0, 1]$. This ensures that the personal dissatisfactions of the members of the out-group are counted as less important in the aggregation than the personal dissatisfactions of the members of the in-group.[14] We then apply the weight when computing aggregate dissatisfaction across individuals to generate the norm function for agents in group N_1:

$$D_{N_1}(x) := \sum_{i \in N_1} D_i(x) + \omega \sum_{i \in N_2} D_i(x). \tag{1}$$

Notice that this is the aggregate dissatisfaction of players in group N_1. In general, it will be different from a similar object written for N_2 since there the members of N_1 would be discounted instead.[15]

We analyze the study by Chen and Li (2009) 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 5. 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' own payoffs are unaffected by these decisions). Figure 8 shows the results.

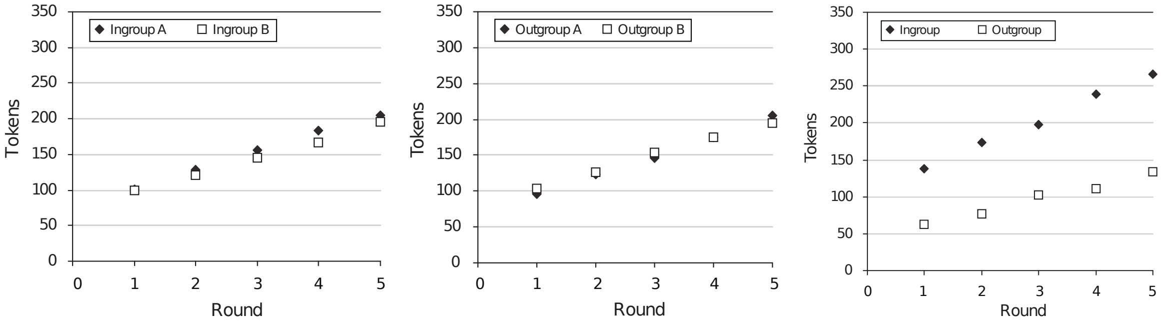


Figure 8: Choices in the other-other allocation tasks of Chen and Li (2009) with different compositions of others.

[14] In principle, ω could even be negative; this would legitimize outright hostility towards the out-group.

[15] 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.

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 $\omega = \frac{1}{2}$, which implies that subjects treat the dissatisfaction of in-group members as twice as important as that of out-group members. We can use this weight to test the comparative statics of the theory of injunctive norms using data from the second part of the experiment.

After this task subjects played in a sequence of 23 one- or two-moves games drawn 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.3. The norm functions are computed with the weight $\omega = \frac{1}{2}$ 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 (see the last two columns of Table 2). The former is the difference in the proportion of players B (second movers) choosing Left 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 by player B.[16] This quantity measures the change in norm-dependent utility of the second mover when playing the game with an in-group member vs. with an out-group member.

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 9 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. □

[16] 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; see Kimbrough and Vostroknutov (2023b) for a discussion of norm-driven punishment. 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.

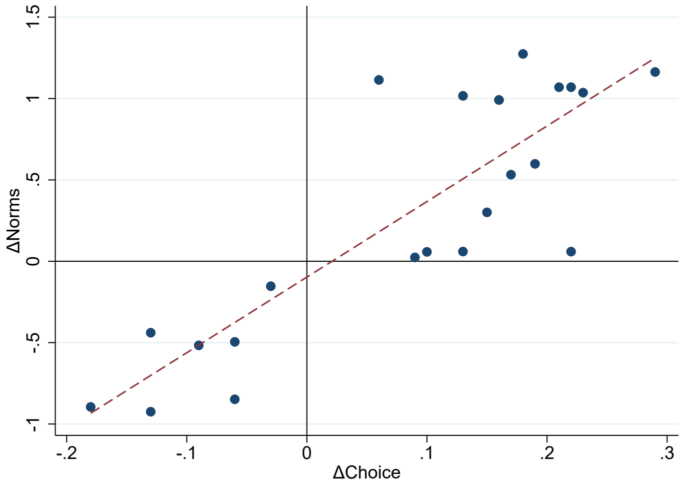


Figure 9: Change in choices and change in normative valences predicted by our model for the 23 games studied in Chen and Li (2009).

4 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 power in explaining other-regarding and context-dependent behavior. 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 the normative appropriateness of each outcome can be defined only in terms of the set of possible outcomes in a game.

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 fact that in- and out-groups are treated differently.

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.

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. 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.

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. For example, whether a non-vegan diet is normatively appropriate depends on whether (and how much) we count animals in our normative calculus. Similarly, we assume an implausibly powerful ability to empathize with others, requiring complete knowledge of others' preferences over outcomes to get norms off the ground. This highlights another important source of normative disagreement: lacking knowledge of others' preferences 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). Moreover, to account for behavioral context-dependence on seemingly "irrelevant" alternatives, we assume that dissatisfactions depend on the entire set of feasible consequences. However, the larger the set of consequences, the less likely people are to be capable of fully considering them all. This suggests extending our framework to incorporate notions of focality or salience of particular consequences in order to better understand how the presentation of information about the set of consequences can shape normative judgment. If a reader has come with us this far, then we hope that these complications will be a wellspring of future research.

References

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, 87(4):1115–1153.

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.

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

Battigalli, P., Corrao, R., and Dufwenberg, M. (2019a). Incorporating belief-dependent motivation in games. *Journal of Economic Behavior & Organization*, 167:185–218.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2):170–176.

Battigalli, P., Dufwenberg, M., and Smith, A. (2019b). Frustration, aggression, and anger in leader-follower games. *Games and Economic Behavior*, 117:15–39.

Bénabou, R. and Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, 126(2):805–855.

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

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10:122–142.

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.

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.

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

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

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3):817–868.

Fioretti, M., Vostroknutov, A., and Coricelli, G. (2022). Dynamic regret avoidance. *American Economic Journal: Microeconomics*, 14(1):70–93.

Galeotti, F., Montero, M., and Poulsen, A. (2019). Efficiency versus equality in bargaining. *Journal of the European Economic Association*, 17(6):1941–1970.

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

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. (2023a). A meta-theory of moral rules. mimeo, Chapman University and Maastricht University.

Kimbrough, E. O. (2022). Rules, perception and the intelligibility of laboratory experiments on social interaction in economics. In *Contemporary Methods and Austrian Economics*, volume 26, pages 35–53. Emerald Publishing Limited.

Kimbrough, E. O. and Vostroknutov, A. (2023b). Resentment and punishment. Mimeo.

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.

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

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.

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.

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.

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.

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

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.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211:453–458.

Vostroknutov, A. (2020). Social norms in experimental economics: Towards a unified theory of normative decision making. *Analyse & Kritik*, 42(1):3–40.

Appendix (for online publication)

A Additional Supporting Evidence

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

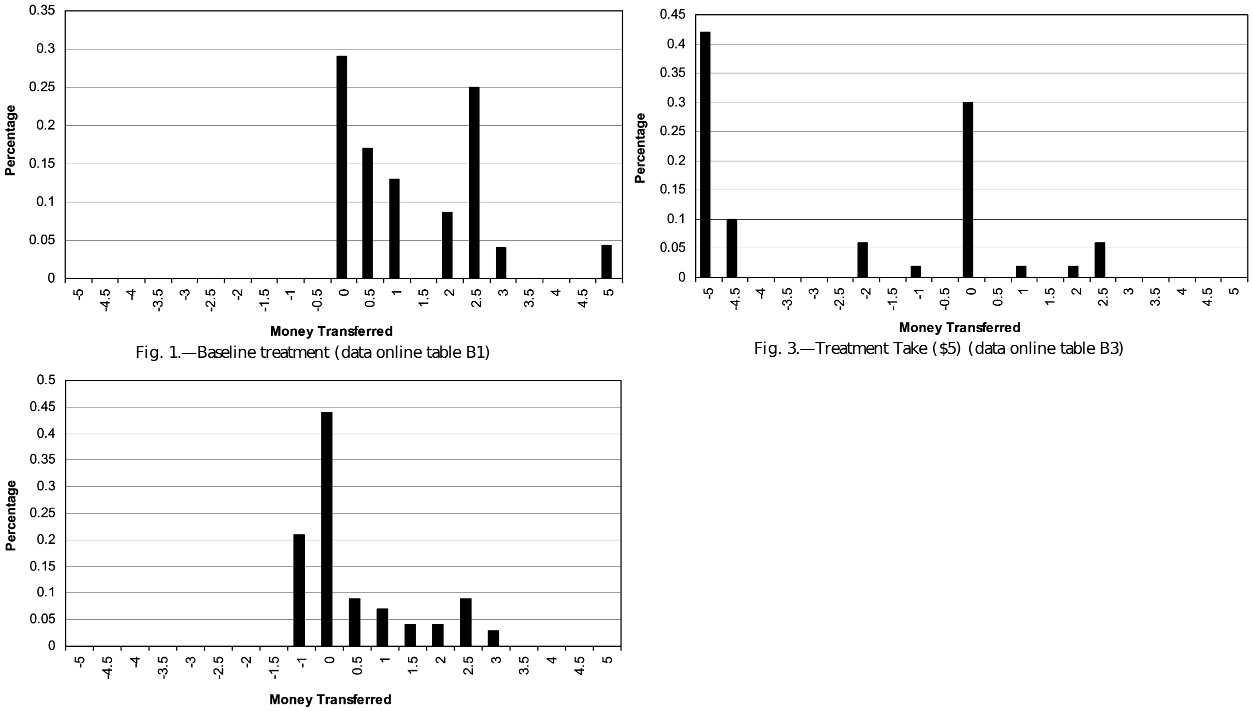


Fig. 1.—Baseline treatment (data online table B1)

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

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

Figure 10: Data from List (2007).

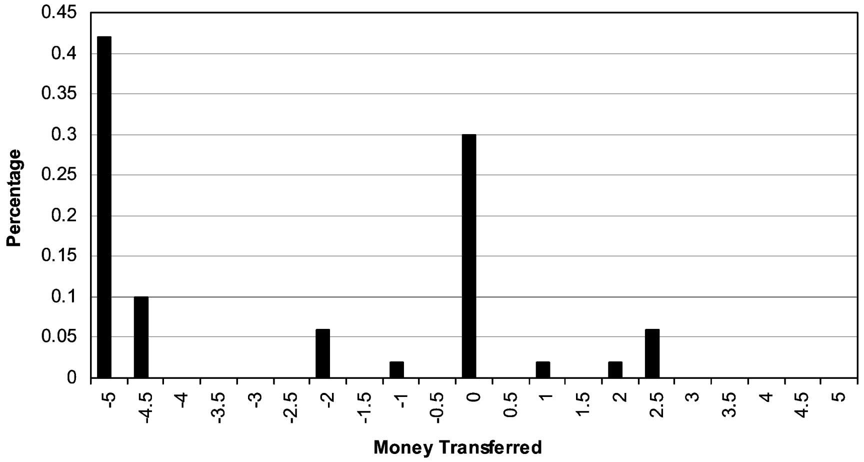


Fig. 3.—Treatment Take ($5) (data online table B3)

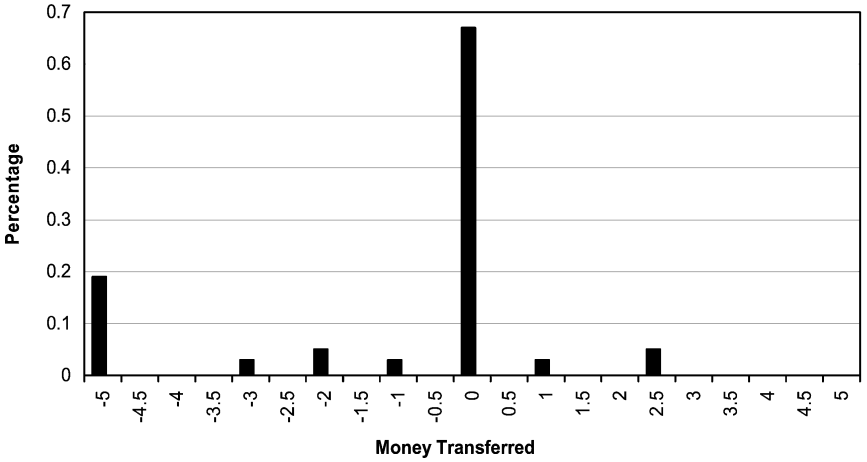


Fig. 4.—Treatment earnings (data online table B4)

Figure 11: Data from List (2007).

A.2 Supporting Evidence for Case 4. McCabe et al. (2003).

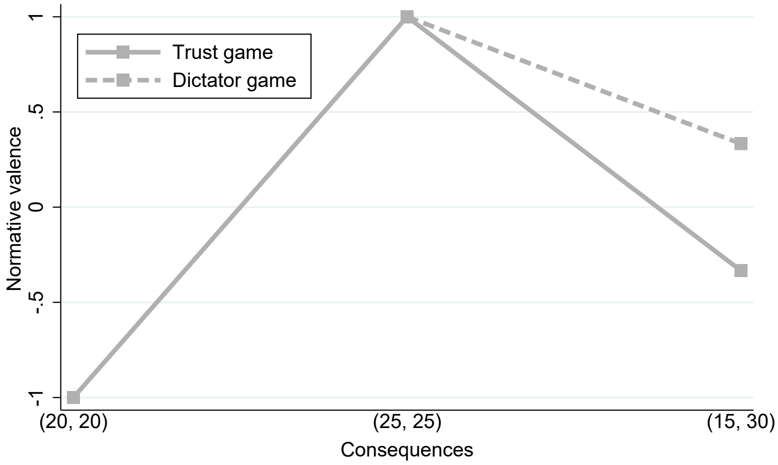


Figure 12: Norm functions with linear consumption value functions for the games analyzed in McCabe *et al.* (2003).

A.3 Supporting Evidence for Case 5. 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 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 need a way to compare normative valences of a consequence x that belongs to two different sets of consequences C_1 and C_2 ($x \in C_1 \cap C_2$). We assume that the payoffs from x, $u(x)$, 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 $\eta(x|C_1)$ and $\eta(x|C_2)$. In order to compare these norm functions on $C_1 \cap C_2$, we need to find some common ground, since $\eta(x|C_1)$ and $\eta(x|C_2)$ are normalized using completely different dissatisfactions. We postulate that if a consequence $x_i \in C_i$, $i \in \{1, 2\}$, is the most appropriate in its corresponding set or $\eta(x_i|C_i) = 1$, 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 $m_i = \min_{x \in C_i} D^i(x)$ and $m = \min_{i \in \{1,2\}} m_i$, and redefine the dissatisfactions as $\bar{D}^i(x) = D^i(x) - m_i + m$, so that the lowest dissatisfaction (for the most appropriate consequence) is the same in both environments.[1] Let $w = \max_{i \in \{1,2\}} \max_{x \in C_i} \bar{D}^i(x)$ be the highest dissatisfaction in all environments and use the interval $[m, w]$ 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 $\ddot{\eta}(x|C_i) := [-\bar{D}^i(x)]_{[-w,-m]}$ 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 $\ddot{\eta}(x|C_i)$ is calculated as $-\bar{D}^1$, which is normalized from the interval that covers dissatisfactions in both environments to $[-1, 1]$.

[1] Note that adding a constant to D^i or multiplying it by a positive constant does not change the associated norm function $\eta(x|C_i)$.

C Proofs

Proof of Proposition 1. $(1 \Rightarrow 2)$. By A1 $d_i(t,r) = d_i(0,r-t)$. When $r-t > 0$, by A2 it is true that $d_i(0,r-t) = (r-t)d_i(0,1)$, and by A3, whenever $r-t \leq 0$ we have $d_i(0,r-t) = 0$. By A4 then $d_i(0,r-t) = r-t$. Thus, we can write $d_i(t,r) = \max\{r-t,0\}$. △

$(2 \Rightarrow 1)$. A1-A4 hold trivially. □

Proof of Proposition 2. $(1 \Rightarrow 2)$. Take any finite C with more than one element and take any $x \in C$. Enumerate the elements of C:

$$C = \{x_1, x_2, ..., x_K\} \cup \{x\}.$$

By A5 $D_i(x \,|\, \{x\}) = 0$ and by A6 $D_i(x \,|\, \{x, x_1\}) = d_i(u_i(x), u_i(x_1))$. Add elements one by one and use A6 repeatedly to get

$$D_i(x \,|\, C) = \sum_{j=1}^{K} d_i(u_i(x), u_i(x_j)) = \sum_{c \in C \setminus \{x\}} d_i(u_i(x), u_i(c))$$

as desired. △

$(2 \Rightarrow 1)$. A5 and A6 hold trivially. □

Proof of Proposition 3. $(1 \Rightarrow 2)$. For all $D_1, ..., D_N \in \mathbb{R}_+$ A3 implies $G(D_1, ..., D_N) = G(0, ..., 0) + \sum_{i \in N} \omega_i D_i$. By A1, $G(D_1, ..., D_N) = \sum_{i \in N} \omega_i D_i$. Thus, since $D_i(x \,|\, C)$ satisfy A5-A6 and d_i satisfy A1-A4, we have

$$D(x \,|\, C) = \sum_{i=1}^{N} \omega_i D_i(x \,|\, C) = \sum_{i=1}^{N} \sum_{c \in C} \omega_i \max\{u_i(c) - u_i(x), 0\}$$

as desired. △

$(2 \Rightarrow 1)$. A7-A8 are trivial. We get A1-A6 from the proofs of Propositions 1 and 2. □

Proof of Proposition 4. Consider a finite context $C \subset \mathcal{C}$ and two consequences $x, y \in C$ with $u_i(x) \geq u_i(y)$ for all $i \in N$ with at least one strict inequality. For any i with $u_i(x) = u_i(y)$ we have $D_i(x|C) = D_i(y|C)$, and for any i with $u_i(x) > u_i(y)$ it is true that

$$D_i(x|C) = \sum_{z \in C} \max\{u_i(z) - u_i(x), 0\} < \sum_{z \in C} \max\{u_i(z) - u_i(y), 0\} = D_i(y|C).$$

Thus, $D(x|C) < D(y|C)$. The inequality is strict since $d_i(y,x) > 0$ for i with $u_i(x) > u_i(y)$. □

Proof of Proposition 6. For any consequence c_j the aggregate dissatisfaction is given by

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

which can be rewritten as

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

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

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

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. □

Proof of Proposition 7. Let us begin with calculating the normative value of (z, z). The points (a_i, b_i) are irrelevant for this since they are Pareto-dominated by (z, z) or equal to it. Thus, they do not evoke dissatisfaction at (z, z). The pairs of points (x_i, y_i) and (y_i, x_i) only influence dissatisfaction of (z, z) through y_i's and not x_i's since they are less than or equal to z. Therefore, the dissatisfaction at (z, z) is

$$D(z, z) = 2\sum_{i=1}^{n}(y_i - z),$$

which is shown in Figure 13.

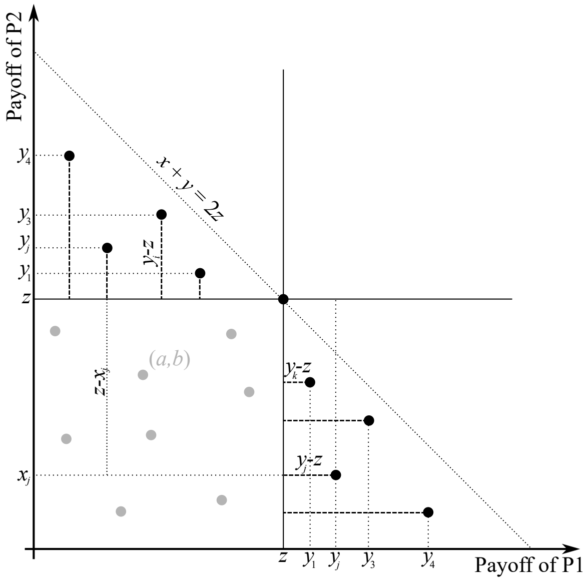


Figure 13: Illustration of the dissatisfaction calculations.

Now, fix any index j and consider the point (y_j, x_j). The dissatisfaction $D(y_j, x_j)$ can be written as

$$D(y_j, x_j) = \sum_{i=1}^{n}(y_i - z) + (n+1)(z - x_j) + \sum_{i=1}^{n}(y_i - z) - \sum_{k<j}(y_k - z) - \sum_{k>j}(y_j - z) + \delta_j.$$

Here $(n+1)(z - x_j)$ is the additional dissatisfaction as compared to $D(z, z)$ because of the points (x_i, y_i) and (z, z), the two sums with k is the additional dissatisfaction because of the points (y_i, x_i), and $\delta_j \geq 0$ is the dissatisfaction because of points (a_i, b_i) and the points (y_i, x_i) with $x_i \geq x_j$. Figure 13 illustrates. Thus, the difference in dissatisfactions between $D(y_j, x_j)$ and $D(z, z)$ is equal to

$$\Delta = (n+1)(z - x_j) - \sum_{k<j}(y_k - z) - \sum_{k>j}(y_j - z) + \delta_j = (n+1)(z - x_j) - \sum_{k<j}(y_k - z) - (n - j)(y_j - z) + \delta_j.$$

Using the assumed condition $x_i + y_i \leq 2z$, which is the same as $z - x_j \geq y_j - z$, we get

$$\Delta \geq (n+1)(y_j - z) - \sum_{k<j}(y_k - z) - (n-j)(y_j - z) + \delta_j$$

or

$$\Delta \geq (j+1)(y_j - z) - \sum_{k<j}(y_k - z) + \delta_j.$$

This can be rewritten as

$$\Delta \geq 2(y_j - z) + \sum_{k<j}(y_j - y_k) + \delta_j \geq 0.$$

The last inequality follows from the assumption that $z \leq y_1 \leq y_2 \leq ... \leq y_n$. Therefore, the dissatisfaction of any point (y_j, x_j) is weakly higher than that of $D(z, z)$. Points (a_i, b_i) also have higher dissatisfaction than (z, z) because they are Pareto-dominated by it (see Proposition 4). This makes (z, z) the norm. □

Proof of Proposition 8. Consider the payoff vector $(z, z) = (p_1^* + p_2^*, p_1^* + p_2^*)$. All gift exchange games can be divided into two classes depending on the Pareto-dominance properties of (z, z). In the first class (z, z) Pareto-dominates all other possible payoff vectors. In this case (z, z) is the norm by Proposition 4 and has zero dissatisfaction. In the second class (z, z) is Pareto-optimal but there are other payoff vectors that are on the Pareto frontier. Let us call the collection of pairs of such vectors (x_i, y_i) and (y_i, x_i) for $i = 1..n$, where n is the number of such pairs. They are symmetric around the 45° line by the property of the gift exchange game, and without loss of generality satisfy the conditions $x_i \leq z$ for all $i = 1..n$ and $z \leq y_1 \leq y_2 \leq ... \leq y_n$ since they are on the Pareto frontier. Finally, any such point (x_i, y_i) can be written in terms of game payoffs as $(p_{1t} + p_{2r}, p_{2t} + p_{1r})$ for some $t, r \in E$. Notice that by the definition of (z, z) we have $p_{1t} + p_{2r} + p_{2t} + p_{1r} \leq 2z$, or $x_i + y_i \leq 2z$ for all $i = 1..n$. Thus, all the conditions of Proposition 7 are satisfied, which implies that (z, z) is the norm. □

Proof of Proposition 9. Let $\langle N, C, u \rangle$ be a tournament. Set C consists of $N!$ consequences corresponding to all possible assignments of the prizes, and in each consequence each payoff from $\{x_1, x_2, ..., x_N\}$ happens only once. Assume without loss of generality that $x_1 \leq x_2 \leq ... \leq x_N$. If we look at the payoffs of player i in all consequences, we find that she receives any payoff x_j in $(N-1)!$ consequences. Slightly abusing notation, we can express the dissatisfaction of the consequence that gives player i payoff x_j as

$$D_i(x_j) = (N-1)! \sum_{\ell=j+1}^{N} (x_\ell - x_j).$$

This amount is the same for all players. Since in each consequence each payoff happens exactly once, the aggregate dissatisfaction is the same for each consequence. □

D Compromise Theorem

In this appendix we show the technique with which general results can be proven for the minima of the aggregate dissatisfaction functions generated by the model. Our primary interest here is to find some properties of the norm (the consequence that gives the decision-maker the lowest aggregate dissatisfaction) that hold on some relatively large class of contexts. The reason we are interested in this is that there are many games with continuous action spaces (e.g., Trust, Public Goods, Common Pool Resource games) that can provide valuable intuition into moral behavior. However, within the model, it can be difficult to deal with such games due to the following. First, the axioms deal with *finite* sets of consequences. Thus, we want to know the properties of the norm in games like Public Goods for arbitrarily precise discretizations of their action spaces. For this we will use Lebesgue integration on the convex hulls of sets of allocations in $\mathbb{R}^N$. Second, such integration can be problematic given that the sets of allocations are convex subsets of $\mathbb{R}^N$ that do not have a product-space property (they cannot be written as product spaces and integrated sequentially by each dimension).

For games with N players, we consider a finite set of consequences C, the image of the consumption value vector $u[C]$ in $\mathbb{R}^N$ and its corresponding convex hull Ω, which is a convex N*-polytope*.[2] A convex polytope in $\mathbb{R}^N$ is a set $\{x \in \mathbb{R}^N \mid Ax \leq b\}$ defined by a collection of m linear inequalities $Ax \leq b$, where A is an (m, N)-real matrix and $b \in \mathbb{R}^m$. We focus on polytopes because the sets of allocations in many games, like the Trust game or the Public Goods game, are convex polytopes (see Figure 3 for the examples in $\mathbb{R}^2$). Notice several things about this definition. First, a convex polytope can be alternatively seen as a convex hull of its *vertices* in $\mathbb{R}^N$. Second, N-polytope is a subset of $\mathbb{R}^N$ that is bounded by $(N-1)$-dimensional faces, which in their turn are $(N-1)$-polytopes. Therefore, these $(N-1)$-faces are bounded by $(N-2)$-polytopes, etc. As we reduce dimensionality in this way, we converge to 1-faces, or *edges*, that connect vertices to each other. Finally, if we have a convex N-polytope and we cut it into two parts by an $(N-1)$-dimensional hyperplane, we get two N-polytopes that together constitute the original one. This is an important property that we will use in what follows.

D.1 Integration over Polytopes

When Ω is an N-polytope, the computation of the aggregate dissatisfaction function at some point $x \in \Omega$ involves integration of dissatisfactions of player i (some linear function) over the sub-polytope defined by the intersection of Ω with the subspace $T_{x_i} = \{y \in \mathbb{R}^N \mid y_i \geq x_i\}$, where x_i is the ith component of x. This is necessitated by the fact that dissatisfactions of i at x are positive only for allocations that give i more than x_i and are zero otherwise. Thus, in order to understand the shape and the properties of the aggregate dissatisfaction function on Ω, we need to understand how to integrate linear functions on polytopes.

To do that, we use the result of Lasserre (1998) that reduces the integration of a continuous function $f : \Omega \to \mathbb{R}$ homogenous of degree 1 to the sum of integrals over Ω's $(N-1)$-faces denoted by Ω_k^{N-1} where k enumerates all m such faces (each is defined by one inequality from $Ax \leq b$). The formula for this is given in Theorem 2.4 of Lasserre (1998):

$$\int_{\Omega} f(x)dx = \frac{1}{N+1} \sum_{k=1}^{m} \frac{b_k}{\|A_k\|} \int_{\Omega_k^{N-1}} f d\mu.$$

Here b_k is the kth component of b, $\|A_k\|$ is the Euclidean distance of the kth row of A (as an N-vector) from the origin, and μ denotes the $(N-1)$-dimensional Lebesgue measure on $(N-1)$-polytope Ω_k^{N-1}.

[2] Several remarks are in order at this point. First, we assume that the mapping $C \mapsto u[C]$ is a bijection. In other words, each consequence in C is mapped by u into a unique allocation in $\mathbb{R}^N$ (this is true in the Trust and Public Goods games, as well as many others). Second, we assume that Ω is N-dimensional (has non-zero N-dimensional Lebesgue measure). And third, notice that since C is finite, Ω is compact.

For the purpose of discerning the properties of the aggregate dissatisfaction function, we iterate this formula by recursively applying it first to all Ω_k^{N-1} and then to consequent polytopes of lower dimensions until we reach the edges of Ω denoted by Ω_k^1. As a result, we can obtain the following:

$$\int_\Omega f(x)dx = \sum_{k=1}^{m'} \xi_k \int_{\Omega_k^1} f d\mu.$$

Here the sum goes over all m' edges of Ω and $\xi_k \in \mathbb{R}$ are coefficients obtained from the recursion (some combinations of N and entries in A and b). The important thing to notice about this formula is that we can reduce the integration of f over N-polytope Ω to the integration over its edges Ω_k^1. The integration of each Ω_k^1 is straightforward to do, since it is simply an integral of f over some interval in $\mathbb{R}$. This result will allow us to characterize the norm function in certain conditions without knowing what ξ_k exactly are.

D.2 Personal Dissatisfaction Function

With these results in mind, we now express the personal dissatisfaction of player i in terms of integrals over polytopes. According to Proposition 3, the axioms are equivalent to i's dissatisfaction function

$$D_i(x \mid C) = \sum_{c \in C} \max\{u_i(c) - u_i(x), 0\}$$

on finite sets C. Notice that we defined our axioms only on the finite sets C, which in principle precludes the usage of full-dimensional subsets of $\mathbb{R}^N$ and integration. However, for the kind of games that we have in mind (e.g., Trust, Public Goods), the continuum can be thought of as an approximation of the arbitrarily precise but finite subsets of allocations in these games that comes from the fact that money is not infinitely divisible (up to cents, for example). Thus, even though we did not formally define our axioms on continuous sets C (the work for the future research), we nonetheless will use the continuous formulation of the personal dissatisfaction function D_i on Ω to understand its properties. The personal dissatisfaction of i in allocation x (as an approximation of discrete cases) can be expressed as

$$D_i(x \mid \Omega) = \int_\Omega \max\{u_i - x_i, 0\} du.$$

Here the variable of integration u goes over the set Ω and u_i corresponds to its ith component. This formulation is equivalent to

$$D_i(x \mid \Omega) = \int_{\Omega(x_i)} (u_i - x_i) du,$$

where $\Omega(x_i) = \Omega \cap T_{x_i}$ is the original polytope restricted to the allocations that give player i at least x_i. Thus, $D_i(x \mid \Omega)$ is an integral of a linear function over the polytope $\Omega(x_i)$. Using the result in the previous section, we can express this in the following way:

$$D_i(x \mid \Omega) = \sum_{k=1}^{m'} \xi_k \int_{\Omega_k^1(x_i)} (u_i - x_i) du,$$

where the sum goes over all edges $\Omega_k^1(x_i)$ of $\Omega(x_i)$. Notice that our original problem is now reduced to computing the integrals of linear functions over 1-dimensional edges, which boils down to areas of triangles and rectangles in two dimensions.[3]

[3] We abuse notation slightly by using the non-homogenous function $u_i - x_i$ of u. However, $u_i - x_i$ becomes homogenous if we just subtract x_i from the ith component of each point in $\Omega(x_i)$. Thus, the two formulations are equivalent.

D.3 The Scarcity Condition

The main goal of this appendix is to demonstrate that we can translate general geometric properties of the sets of allocations represented by convex polytopes into the properties of aggregate dissatisfaction functions coming from the P-axioms. Here we define one geometric condition that can be translated into some useful property of the minimum of the aggregate dissatisfaction function.

Many social dilemmas with continuous action sets, like the Dictator, Trust, or Public Goods games, have an interesting property that is rarely discussed in the literature. Namely, in these games it is *not* the case that several players can achieve their highest possible payoffs at one allocation. Rather to the contrary, one player can achieve her maximal payoff only when another player sacrifices a lot of his payoff. In the Trust game (see Figure 3) the second mover should give all his money to the first mover for her to achieve the highest payoff (or give the first mover nothing to achieve his highest payoff). Geometrically, this property means that there is only one point (the vertex of the polytope) where the highest payoff of each player is achieved.

As will become clear below, our axioms suggest that this observation might be expressing a property that any game should satisfy for it to be (intuitively) counted as a social dilemma. Thus, we generalize the property to any convex polytope. Consider any N-polytope Ω, fix some player i, and suppose that her highest achievable consumption value in Ω is z_i. Then, Figure 14 shows two possibilities that can arise with regard to the number of allocations where player i gets z_i. Either it can be one allocation as on the left panel of Figure 14, or it can be a continuum of allocations (the right panel). There are no intermediate cases since Ω is convex.

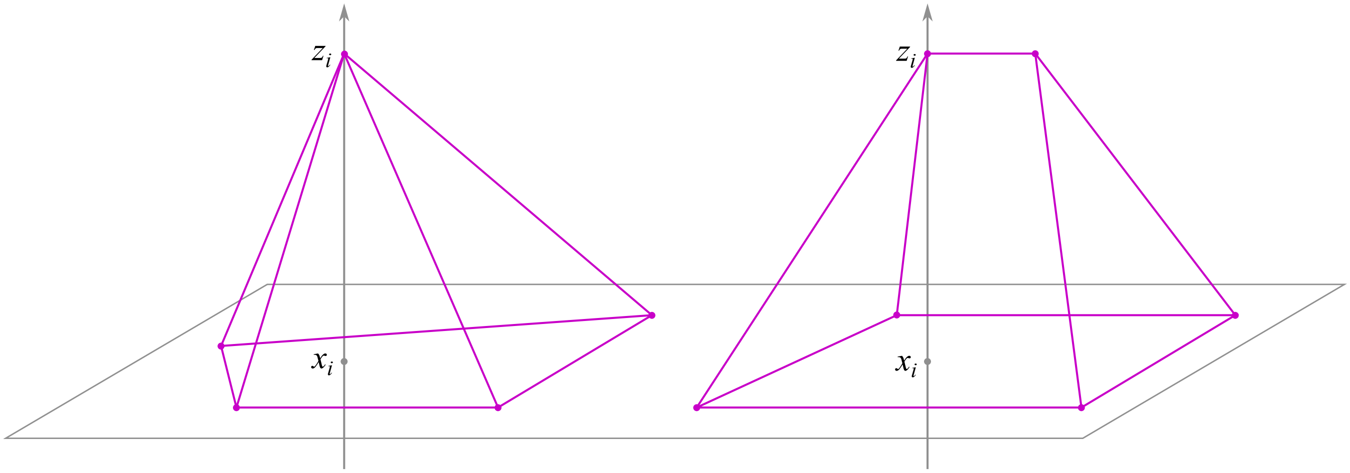


Figure 14: **Left:** The case when player i's highest consumption value z_i is achieved in only one allocation. **Right:** The case when player i's highest consumption value z_i is achieved in a continuum of allocations.

If z_i is achieved in only one allocation, then it is clear that there is high enough level of i's consumption value, say x_i, at which all allocations that give i consumption value x_i or more form a pyramid as on the left panel of Figure 14 (the pyramid is formed by the edges going from the i's iso-value plane containing x_i to the allocation that gives i consumption value z_i). This fact follows from the property of Ω that is a convex hull of a *finite* number of vertices. Similarly for the right panel: there is some x_i such that the subpolytope with consumption value of player i higher or equal to x_i is formed by edges that go from the

i's iso-value plane containing x_i to some allocation that gives i consumption value z_i. With this in mind, we define the Scarcity Condition for convex N-polytopes.

The Scarcity Condition *Suppose that Ω is a convex N-polytope. Then we say that Ω satisfies the Scarcity Condition if for each $i \in N$ there is only one allocation $Z_i \in \Omega$ where the maximum consumption value z_i of i is achieved. Moreover, all allocations Z_i are different for all i. In other words, $\forall i, j \in N \;\; Z_i \neq Z_j$.*

In its essence, the Scarcity Condition states that if player i achieves the highest consumption in some allocation Z_i, then no other player can enjoy the highest value at the same time. This expresses an idea that the consumption value in the game defined by Ω is somehow scarce for otherwise it would be possible to have more than one player enjoying the highest value at once. The uniqueness of Z_i also expresses a version of scarcity in the sense that if Z_i is achieved, then any redistribution of value within Ω will lead to i getting strictly less of it. In the next section we show what this condition implies for the dissatisfaction of player i.

D.4 Properties of Personal Dissatisfaction Functions

The goal of this section is to understand how the personal dissatisfaction function of player i behaves at allocations that give i the maximal consumption value z_i. We need this to prove our main result in the next section.

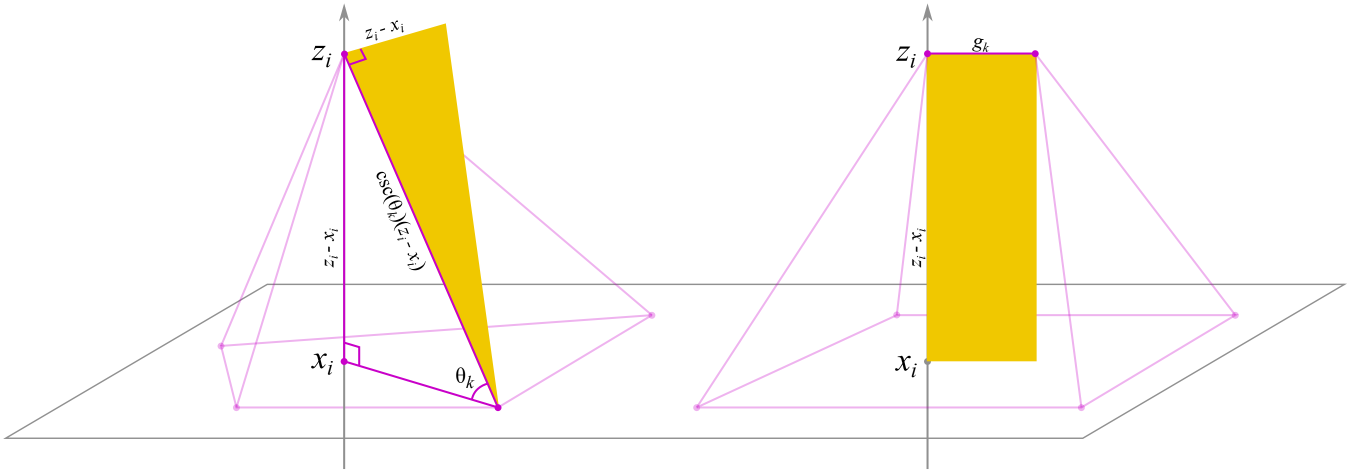


Figure 15: **Left:** A computation of the integral along the edge of the pyramid. **Right:** A computation of the integral along the horizontal edge with maximal consumption value z_i.

As we know from Section D.2, the personal dissatisfaction at allocation x that gives i consumption value x_i is the weighted sum of integrals of linear functions along the edges $\Omega_k^1(x_i)$ of the polytope $\Omega(x_i)$. Let us apply this result to the pyramid on the left panel of Figure 15 that is a typical situation for any polytope Ω that satisfies the Scarcity Condition. The figure shows the computation of the integral for one edge k which is at angle θ_k to the i's iso-value plane containing x_i. Given this, the length of the edge is $\csc(\theta_k)(z_i - x_i)$, where $\csc(\theta_k) = 1/\sin(\theta_k)$ is the cosecant. The integrand linear function on this edge changes from 0 to $z_i - x_i$. Thus, the integral is equal to the area of the right triangle (in yellow) with catheti equal to $\csc(\theta_k)(z_i - x_i)$ and $z_i - x_i$, which gives us

$$\int_{\Omega_k^1(x_i)} (u_i - x_i) du = \frac{\csc(\theta_k)}{2}(z_i - x_i)^2 = \nu_k (z_i - x_i)^2,$$

with $\nu_k = \frac{\csc(\theta_k)}{2}$. This is true for all k representing edges of the pyramid that end at z_i. For the edges that lie on the x_i-iso-value plane, the integrals are zero since the integrand function is zero. Therefore, this gives us the exact formula for the value of the personal dissatisfaction function of i in the vicinity of the allocation that gives i maximal consumption value z_i:

$$D_i(x \mid \Omega) = \sum_{k=1}^{m''} \xi_k \nu_k (z_i - x_i)^2 = \pi_i (z_i - x_i)^2,$$

where $k = 1$ to m'' enumerates all edges that end at z_i and $\pi_i = \sum_{k=1}^{m''} \xi_k \nu_k$. This is a simple parabola that reaches zero at point $x_i = z_i$. More importantly, the derivative of $\pi_i(z_i - x_i)^2$ with respect to x_i is also zero at $x_i = z_i$.

Notice also another important property of $D_i(x \mid \Omega)$. Let us replace x_i with the variable $y_i = z_i - x_i$. Then, it is clear that $D_i(y \mid \Omega)$ is an increasing function of y_i. This is so for the following reason. Consider any y_i and $y_i + \varepsilon$, where $\varepsilon > 0$ is an arbitrarily small number (going away from the maximum consumption value point z_i). Then, the personal dissatisfaction of i at $y_i + \varepsilon$ is larger than that at y_i because 1) i is dissatisfied about allocations in $\Omega(y_i)$ even more at $y_i + \varepsilon$ than at y_i; 2) additional allocations in between y_i and $y_i + \varepsilon$ also create dissatisfaction. In addition, as more and more edges are added into the calculation of $D_i(y \mid \Omega)$ as y_i increases, we get the function that is constructed piece-wise from quadratic forms and is increasingly convex as more and more edges are included. We formulate this as a proposition.

Proposition 10. *When Ω satisfies the Scarcity Condition, the personal dissatisfaction function $D_i(y \mid \Omega)$ as a function of distance $y_i = z_i - x_i$ from z_i is an increasing piece-wise convex parabola with minimum at the allocation that gives i the maximal consumption value z_i. Moreover, the derivative of $D_i(y \mid \Omega)$ at z_i is zero. This is true for all i.*

Proof See the argument above.

Notice the importance of the Scarcity Condition for this result. Indeed, if we look at the right panel of Figure 15, where the Scarcity Condition is not satisfied, we get something rather different. While we can compute the integrals for each edge that goes from the x_i-iso-value plane to z_i in the same way as above, the integral along the top edge k of length g_k (where $x_i = z_i$ for all allocations) is equal to the area of the yellow rectangle, namely $g_k(z_i - x_i)$. Thus, the personal dissatisfaction of i at x_i will be an expression of the type $\pi_i(z_i - x_i)^2 + \xi_k g_k (z_i - x_i)$, where the first term is as above and the second term represents the area of the rectangle times some positive constant. The derivative of this expression at $x_i = z_i$ is not zero and is equal to $-\xi_k g_k$. This fact will be used to prove the main result in the following section.

D.5 The Compromise Principle

With all the previous findings, we are finally ready to formulate our main result in this appendix. We claim that the minimum of the aggregate dissatisfaction function

$$D(x \mid \Omega) = \sum_{i=1}^{N} \omega_i D_i(x \mid \Omega)$$

on any convex N-polytope Ω that satisfies the Scarcity Condition is *not* at the allocations that give any one player the maximal consumption value. In such situations allocating any player the maximal value is not the normatively best thing to do. In other words, compromises always need to be made. We state this as a separate definition.

The Compromise Principle *Allocating any one player the maximal possible consumption value is not the normatively best thing to do.*

Showing that the normatively best allocation is never at the vertices of Ω where one player gets the maximal consumption value is rather straightforward given the findings above. Indeed, take the allocation Z_i where i gets the maximal value. Then, the derivative of i's personal dissatisfaction at that point is zero (since Ω satisfies the Scarcity Condition). However, the derivatives of personal dissatisfactions of all other players are positive in Z_i (in some direction) because by Proposition 10 these are increasing piecewise parabolae with zero derivatives not in Z_i but elsewhere (also by the Scarcity Condition). This means that at Z_i the derivative of the weighted sum of personal dissatisfactions (which is $D(x \mid \Omega)$) is positive in some direction. This shows that points Z_i cannot be the minima of $D(x \mid \Omega)$. We formulate this as a proposition that we call the Compromise Theorem to emphasize that in the normatively best outcome no player gets maximal consumption value.

Proposition 11. (Compromise Theorem) *When Ω satisfies the Scarcity Condition, the minimum of the aggregate dissatisfaction function $D(x \mid \Omega)$ follows the Compromise Principle.*

Proof See the argument above.

Notice that the Scarcity Condition is crucial for this result, as if we have a situation that violates it (as in the right panel of Figure 15), then the derivative of D_i at Z_i will be negative and it is possible then to have a situation where the normatively best allocation is the one where i gets the maximal consumption value. This is because the influence of the dissatisfactions of other players can always be made small enough (by manipulating their social weights ω_j) so that the negative derivative will overcome the positive derivatives coming from other players.

Two final remarks should be made at this point. First, the Compromise Theorem presented above can be seen as a generalization of Proposition 6, where it is shown that the normatively best outcome (discrete case, two players with constant sum of payoffs) is always the middle allocation. Second, notice that the theorem works for any positive social weights ω_i for $i \in N$. This means that in games that satisfy the Scarcity Condition *everyone* needs to give something up if they want to be norm-following. This refers even to people with arbitrarily high social weights.

Additional References in Appendices

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.

Lasserre, J. (1998). Integration on a convex polytope. *Proceedings of the American Mathematical Society*, 126(8):2433–2441.

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.