A Theory of Moral Reasoning*

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

September 2022

Abstract

We propose a framework for modeling the influence of normative considerations on decision-making and social interaction. Our key insight is that axiomatic models like those used in cooperative game theory can be reinterpreted as theories of moral reasoning and norm formation which describe the principles agents use to reason about the normative properties of a set of feasible outcomes in a choice set or game. We show how to choose axioms to represent different theories of moral reasoning, ranging from simple, abstract moral rules to radically context-dependent moral psychology. Then, under a given theory, we use the axioms to generate an "injunctive norm" which defines an agreed-upon ranking of feasible outcomes from most to least normatively appealing. Paired with a norm-dependent utility specification, this framework provides a powerful tool for modeling the influence of morality on decision-making. We show how the framework relates to the literature on social preferences, and why a theory of moral reasoning rooted in moral psychology is better able to account for the context-dependent nature of decision-making than a simpler theory rooted in moral rules.

JEL classifications: C71, C72, D63, D91
Keywords: norms, axiomatic preferences, rules, morality, decision-making

*We would like to thank Elias Tsakas for help with formulating the axioms and Valerio Capraro, David Rojo-Arjona, Nicole Saito, Matthias Stefan, and seminar audiences at Maastricht University, the University of Konstanz, Virginia Tech, the Max Planck Institute in Bonn, George Mason University, San Jose State University, and The Ohio State University for helpful comments. We thank Maastricht University for funding. 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

In economics, the normative and the positive are traditionally treated separately. This is true for neo-classical theory as well as game theory. Within game theory, one way of thinking about the distinction between cooperative concepts (esp. axiomatic bargaining theory) and non-cooperative ones is that the former provide tools for formalizing particular notions of "the good" and "the bad" which can be used to evaluate the outcomes produced by the (amoral) interactions analyzed by the latter. This separation of spheres, while tidy, overlooks extensive evidence that normative considerations are not merely evaluative but also enter into people's decisions as goals to be aimed at, however imperfectly.

We contend that it is possible to improve our understanding of social decision-making by deepening the integration of cooperative and non-cooperative theory. In this paper, we show how the axiomatic approach, typical for cooperative game theory, can be used to model how people moralize, with different axioms implying different normative rankings of feasible outcomes (hereafter *injunctive norms*). Then, in a companion paper (Kimbrough and Vostroknutov, 2021), we derive the implications of a psychologically plausible set of axioms and use the normative rankings implied thereby as an input to simple models of "norm-dependent preferences," in which decisions reflect trade-offs between normative goals and own-utility maximization. Thus, we show a new way to introduce normative goals into models of choice.

Existing theories in which normative considerations shape choices differ from our approach in that they fail to provide an account of how normative influences 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 moralizing is performed by the modeler, rather than the agents themselves. Social preference theories bake a particular moral goal 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.

In our framework, we directly model how agents engage in moral reasoning by deriving the normative implications of a set of axioms for a given choice set. In this view, axiomatic models drawn from cooperative game theory can be (re)interpreted as models of moral reasoning and norm formation and can then be used to provide a foundation for the kinds of norms that are taken for granted in existing models. Starting from axioms on moral reasoning, which yield individual normative judgments,

we introduce another set of assumptions on how those judgments are aggregated into shared injunctive norms. Then, we show how to take the objects typically analyzed in non-cooperative game theory and construct normative evaluations thereof. A particular set of axioms will imply a particular normative ranking of feasible outcomes. A key consequence is that, for some sets of axioms, an agent's view about what one ought to do can radically depend on the choice context.

As we argue in Kimbrough and Vostroknutov (2021), this is important because a large body of work in behavioral economics reveals spectacular diversity in the criteria that apparently guide human social behavior. Depending on the choice context, existing theory and evidence suggest that people are variously motivated by Pareto improvements, efficiency, equality, maximin, reciprocity, guilt aversion, lying aversion, anger, and so on.[1] Moreover, experimental evidence suggests that simple changes to the choice set are sufficient to cause people to switch from privileging one criterion to another (e.g., List, 2007; Engelmann and Strobel, 2004; Galeotti et al., 2018). Models of norm-dependent preferences have emerged to account for this diversity (López-Pérez, 2008; Cappelen et al., 2007; Kessler and Leider, 2012; Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2016), but the literature so far has failed to provide a coherent account of how and why social norms vary across contexts. Our framework thus offers an attempt to meet that challenge.

In our framework, we first build a model of individual normative judgments and then aggregate those into a collective judgment that accounts for the perspectives of all interested individuals. A key assumption is that individual normative judgments are fundamentally *comparative;* in deciding what one ought to do, one compares the feasible set to some ideal. In particular, we assume that, when evaluating any outcome, a person may realize that they would be *dissatisfied* with that outcome because it compares unfavorably to their ideal. The choice of axioms specifies how one defines this counterfactual ideal. We further assume that an outcome is more dissatisfying the further it is from the ideal. An individual's normative judgment can then be summarized as a function of its distance from the ideal (which need not be a single element of the choice set).

Next, we model aggregation of these individual judgments into a shared norm via a basic human psychological ability: the capacity for empathy, or the ability to imagine how others would feel in similar situations. We assume that all people feel dissatisfaction of the kind described above and that everyone knows that this is true of everyone

[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).

else. We then assume that *norms* emerge that account for the judgments of everyone involved, with the most appropriate outcome being the one that minimizes the aggregated dissatisfaction of interested parties.

A model of moral reasoning, then, defines the nature of dissatisfaction. Once we specify how the ideal is constructed from a choice set, we can define what such a moral reasoner ought (and ought not) do in that setting. Plugging this into a norm-dependent utility specification, we then get a model of the normative tradeoffs faced by such an agent.

To provide intuition, we start by developing a model of agents whose moral reasoning is driven by Nash's axioms, since these axioms are widely known and well-understood. In such a model (and in the right kind of choice context), the most appropriate outcome is that which maximizes the product of agents' consumption utilities. Thus, norm-sensitive agents, motivated by the Nash axioms, may face a tradeoff between maximizing their own consumption utility and choosing allocations which are closer to the Nash Bargaining Solution. Of course, we may have little reason to expect that people's moral reasoning is generally based on the Nash axioms. The point is rather that the choice of axioms is equivalent to specifying a theory of moral reasoning. Crucially, since the axioms are applied to derive a normative ranking endogenously from the available choice set, some degree of context-dependence emerges naturally in our framework.

The extent of context-dependence will depend on the chosen axioms. One attractive approach is to choose axioms that reflect widely known "moral rules" that capture abstract ideal criteria by which all outcomes are evaluated.[2] When derived from a model of this kind, injunctive norms can usually be summarized via general principles like those mentioned above. We show how to use our framework to develop representations that capture the intuition behind popular models in the social preferences literature: for example, we show how to construct models of moral reasoning that result in agents who exhibit forms of inequality aversion, maximin, efficiency-seeking, etc. Thus, we build a bridge between our model and existing approaches to modeling social behavior in the literature.

However, we also show that moral rules based on comparison of outcomes to an ideal reference outcome end up being insufficiently context-dependent. If agents reason according to a single moral rule, then we are unable to account for a number of well-known observations from lab experiments. For example, as noted above, even very subtle changes to the choice set (e.g. adding "irrelevant" nodes to a game that

[2]Our notion of moral rules is most similar to the theory of moral reference points introduced by Cox et al. (2018).

leave the most equal or most efficient outcome unchanged) can lead to changes in behavior. Such radical context-dependence cannot be explained by an abstract moral rule that favors equality or efficiency; nor can evidence that different normative principles seem to drive choices in different circumstances.

We thus introduce our own set of radically choice-set-dependent axioms, which emphasize the psychological bottom-up construction of norms. In our preferred axioms, dissatisfaction is derived from self-interest such that norms vary only with agents' preferences and the choice set. 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 own-utility. Moreover, we assume that agents are dissatisfied not only because of the best alternative outcome, but also because of any feasible outcome that is preferable to the one under consideration. Dissatisfaction with one outcome is the sum over the dissatisfactions induced by all other more preferred outcomes. 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.

Because, in our preferred model, an individual's evaluation of each outcome depends on how it compares to *all* other outcomes, the resulting norms are radically context-dependent. In fact, we prove that the normative evaluation of outcomes implied by our choice-set-dependent axioms cannot be reproduced by any abstract moral rule of the kind introduced above. The implication is that normative reasoning of the kind modeled by our preferred axioms cannot always be summarized by simple, general principles. Nevertheless, in a companion paper (Kimbrough and Vostroknutov, 2021), we show that a functional representation obtained from our preferred axioms offers substantial explanatory power across a wide set of experimental environments in which behavior is known to be context-dependent, with changes in behavior across contexts tracking the predicted changes in norms.

2 Models of Moral Reasoning

Axiomatic models of bargaining and cooperative games often start from a set of normative axioms (e.g., individual rationality, symmetry, IIA) that the "fair bargain" (Nash, 1950)—or the allocation considered the most socially appropriate by all players—should satisfy. As modelers, we tend to think of the axioms as providing the standpoint from which we (modelers) can criticize the outcomes or allocations that arise in non-

cooperative models or in real bargaining scenarios.[3] We might argue that the bargains reached in the equilibria of some class of non-cooperative games violate IIA or generate asymmetric results in symmetric problems, and are therefore normatively undesirable.

While we think this is a valuable use of axiomatic models, we argue that their role can be augmented by a shift in perspective. If we recognize them as models of moral reasoning and norm formation, then we can move the evaluative locus of axiomatic theory from the minds of modelers to the minds of agents themselves. That is, we can use the tools of cooperative game theory and bargaining theory to build models of how *agents* normatively evaluate a set of feasible allocations. Paired with a model of norm-dependent utility (Kessler and Leider, 2012; Krupka and Weber, 2013) in which agents are motivated to do what they believe is normatively good, our framework provides a fuller account of how normative considerations influence choice. The chosen axioms codify the structure of the norms and thus become part of a positive research program aimed at uncovering how norms emerge from agents' cognition, preferences, and constraints.

Existing cooperative solution concepts provide a starting point for thinking about the problem this way. Consider for example, the bargaining problem studied by Nash (1950). Let S be a set of payoff vectors for two players that is compact, convex, symmetric, and includes the origin, which serves as a disagreement point (Nash's assumptions). Such S can represent the set of feasible allocations of some non-cooperative game, say, a Public Goods game with two players. Suppose also that the players have norms entering their (otherwise selfish) utility functions, and that these norms satisfy Nash's axioms. Specifically, the *players* believe that 1) the allocations on the Pareto frontier of S are more socially appropriate than those not on it; 2) the most appropriate allocation satisfies IIA (on subsets of S that include it); and that 3) the most appropriate allocation is symmetric. Using Nash's famous result that, under these conditions, the solution maximizes the product of consumption utilities, we can model the behavior of such norm-abiding agents with a norm-dependent utility function of the form $w_i(x_i, x_{-i}) = x_i + \phi_i \eta(x_1, x_2)$, where $(x_1, x_2) \in S$; x_i represents agent i's linear consumption utility; $\eta(x_1, x_2) = x_1 x_2$ is the *norm function* defined by the axioms; and $\phi_i \geq 0$ is i's general propensity to follow norms (Kimbrough and Vostroknutov, 2016).

Next, we can take any non-cooperative game Γ that has S as the set of feasible allocations and augment it with the norm-dependent utilities w_i, thus obtaining a

[3] The literature on the design of mechanisms to implement various cooperative solutions is probably the clearest illustration of this point since those models take the cooperative solutions as normative and ask how a designer can ensure they are achieved (e.g., Nash, 1953; Perry and Reny, 1994).

modified game $\Gamma'(\phi_1, \phi_2)$ that can be analyzed using any non-cooperative equilibrium concepts. Structured this way, $\Gamma'(\phi_1, \phi_2)$ now represents a positive model of Γ with norm-following agents whose morality is based on the assumptions of the Nash Bargaining Solution (NBS). Notice that $\Gamma'(0, 0) = \Gamma$, so at one extreme, where players do not care about following norms at all, we get the standard non-cooperative game that can be analyzed with some version of Nash equilibrium. At the other extreme, when ϕ_1 and ϕ_2 are very high, the players will choose the NBS in *any* game form that has S as the set of feasible allocations. The latter result follows from the simple fact that, for high enough propensities to follow norms, the players' incentives become exactly aligned, so they both will strive to reach the Nash Bargaining Solution. With measures of ϕ_i in hand that can be obtained from various experimental tasks (e.g., Kimbrough and Vostroknutov, 2018), these predictions are testable.

This example demonstrates our intuition that cooperative solution concepts can be used as falsifiable models of moral reasoning. Our job in developing this line of research is therefore to come up with realistic axiomatizations that overcome the constraints imposed by the known solution concepts. For example, many of them (e.g., NBS, Kalai-Smorodinsky solution) assume that S is convex to guarantee the existence of the solution, or that the solution is Pareto optimal.[4] However, life is full of surprises and non-convexities. Thus, in order to construct useful accounts of morality, we need solution concepts that work on any (compact) set of allocations, that always exist, and that are constructed on the basis of simple psychological assumptions. These requirements make sure that we build axiomatizations that describe plausible accounts of norms that could realistically evolve in our species (Henrich, 2015).[5] A judicious choice of axioms allows us to build models in which norms respond to context and can represent the universe of "as is" human social relationships.

One common feature of social relationships is the frequent reliance on general principles to summarize normative expectations: "waste not", "play fair", "equal effort gets equal reward", etc. Thus, in the next section we introduce a model of moral reasoning in which agents compare outcomes to ideal reference outcomes according to what we call *moral rules*.

[4] A special effort has been exerted in the literature to find equivalent axiomatizations that do not assume Pareto optimality (see e.g., Karos et al., 2018).

[5] Alger and Weibull (2013) also present an account of the emergence of Kantian moral preferences that are evolutionarily stable.

3 Dissatisfaction Functions for Moral Rules

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 some ideal.[6] In this section we describe a set of axioms that incorporate some *moral rule* according to which the dissatisfaction function is constructed.

By moral rule we mean some *abstract* ideal criterion against which all allocations are compared. By abstract, we mean a normative goal defined, in some sense, externally to the game at hand; this goal might refer to concepts like payoff efficiency, Pareto optimality, equality, or some combination of these, or any other general principle that might be offered as a normative guide to behavior. Such moral rules are succinctly summarized, readily codified and learned, and therefore salient both in our daily lives and to researchers who study social behavior. Most of the literature that deals with social welfare and economic efficiency is based on such abstractions. Thus, we start by showing how to represent *any* choice-set-dependent moral rule—which defines the normatively best element in any choice set—with a dissatisfaction function.

We start with a large set $\mathcal{C}$ of all possible consequences, N players, and a utility/payoff function $u : \mathcal{C} \to \mathbb{R}^N$, which for each consequence defines a vector of utilities (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.

Next, we require that for any C there exists a special non-empty set C^*, which represents the set of *ideal* payoff vectors that *in the context of* C are considered the most socially appropriate, or having zero dissatisfaction. In addition, there is a function $u^*(v \mid S)$ that defines an element in S as a "reference point" for v in S. The function exists for each vector $v \in \mathbb{R}^N$ and all subsets $S \subset \mathbb{R}^N$ that are equal to $u[C^*]$ for some C.[7] This function is used to assign to any element of C the ideal element in C^* to which it is compared, or in reference to which the dissatisfaction is expressed. It has the following property: $u^*(v \mid S) = v$ whenever $v \in S$, which makes sure that

[6] Conceptually, these normative evaluations are made prospectively, before an outcome has been realized. We suggest in Kimbrough and Vostroknutov (2021) that an appropriate moniker for the sentiment being captured might be "pre-gret."

[7] Specifically, there is no need to define $u^*(v \mid S)$ for all subsets S of $\mathbb{R}^N$, but only for those that can play a role of a set of ideal payoff vectors $u[C^*]$. This becomes important when we define choice-set-independent dissatisfaction with only one C^* for all C in Section 3.1. There we need to define u^* for only one set S.

the reference point for any ideal element is the element itself. Notice that u^* is only needed when there are sets C^* that are not singletons. If all C^* are singleton sets, as is the case for most ideals, then there is no need to define u^*. Finally, there are functions $f_i : \mathbb{R}^2 \to \mathbb{R}_+$ that define *personal dissatisfactions* for each player $i \in N$. Namely, $f_i(v_i, u_i^*(v \mid S))$ stands for the dissatisfaction that player i feels when her utility is v_i and the utility that she would receive in the ideal situation is $u_i^*(v \mid S)$. This provides us with a theory of individuals' normative evaluations of each outcome as a function of the ideal outcome in the choice set – a choice-set-dependent model of normative judgment.

In the second step we define an *aggregate dissatisfaction function* $F(x \mid C)$ 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. We propose a set of axioms that describe the properties of dissatisfaction f_i, personal dissatisfaction F_i, and aggregate dissatisfaction F. We start with the axioms that define the connection between dissatisfaction f_i and utility of player i, and we refer to these axioms as *R-axioms* (for rule) and the resulting model as the *R-model*.

R1 $\forall i \in N \; \forall t, r, a \in \mathbb{R} \quad f_i(t + a, r + a) = f_i(t, r)$.

R1 states that adding a constant to the utilities 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 utility in all consequences, then her dissatisfaction is unaffected. R1 ensures that d_i can be expressed as a function of the difference of utilities.

R2 $\forall i \in N \; \exists \beta_i \in (0, 1]$ such that $\forall t \geq 0$ we have $f_i(0, -t) = \beta_i f_i(0, t)$.

R2 states that there might be an asymmetry in how dissatisfaction is perceived when the ideal utility is above or below the received utility. Specifically, if player i receives utility 0 when the ideal utility is $-t$, then her dissatisfaction could be lower than the dissatisfaction she gets when the ideal utility is t. Notice that we require $\beta_i > 0$. The reason for this is the following. If β_i is zero, then player i is not dissatisfied when her utility is greater than that in the ideal case. However, then it would be possible that there are consequences for which aggregate dissatisfaction is zero (all players have utility higher than than the ideal one), but which are not in the ideal set. We deliberately exclude such possibilities since by construction C^* should include *all* ideal

elements. So the requirement $\beta_i > 0$ implies that there are no payoff vectors that have zero dissatisfaction but are not included in C^*.

R3 $\forall i \in N \; \forall t, r, \alpha \in \mathbb{R}$ with $\alpha > 0 \quad f_i(\alpha t, \alpha r) = \alpha f_i(t, r)$.

R3 states that if all utilities 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 utilities in a linear way, thus connecting the two concepts. We could have assumed some non-linear, say concave, relationship between dissatisfaction and utility. However, we already allow utility to be a non-linear function of payoffs. R3 reflects an idea that all non-constant marginal effects of payoffs are already encoded in functions u_i.

Finally, we assume non-triviality and importantly equivalence of dissatisfactions across players.

R4 $\forall i \in N \quad f_i(0, 1) = 1$.

R4 serves two purposes. First, it makes sure that players do feel non-zero dissatisfaction. Second, it postulates the "equivalence" of dissatisfactions of all players. This means that all players are equally dissatisfied if they are at a consequence that gives them 0 utils and there exists another consequence that gives them 1 util. This assumption amounts to the claim that the dissatisfaction from the same amounts of utils makes all players dissatisfied in the same way. This is how we operationalize the idea that shared normative evaluations are built on empathy.

The following proposition provides the representation.

Proposition 1 The following two statements are equivalent:

1. f_i satisfies R1-R4*;*
2. $f_i(t, r) = \max\{r - t, 0\} + \beta_i \max\{t - r, 0\}$ for some $\beta_i \in (0, 1]$.

Proof See Appendix A.

The next axiom defines the personal dissatisfaction $F_i(x \,|\, C)$ of player i. This is very straightforward. However, it is worth noting before we state the axiom that we do not require that F_i be zero for singleton choice sets $C = \{x\}$. This is because it is easy to imagine a situation where there is only one allocation which nevertheless deviates from an ideal, for example, when the ideal for each player is the average payoff in x (see Section 3.1). This highlights the sense in which ideals involve abstraction; the ideal may not be among the feasible consequences in C.

R5 $\forall i \in N \; \forall C \subset \mathcal{C}, x \in C \quad F_i(x \,|\, C) = f_i(u_i(x), u_i^*(u(x) \,|\, u[C^*]))$.

R5 says that player i's personal dissatisfaction at consequence x in C is equal to his

dissatisfaction because of an ideal reference element in C^*. This definition looks redundant. However, it creates a very important restriction on the dissatisfactions across different sets of consequences. Specifically, R5 asserts that for any two sets C_1 and C_2 that have $u[C_1^*] = u[C_2^*]$, or have the same ideal payoff vectors, i's personal dissatisfaction of all elements $x \in C_1 \cap C_2$ is the same. In general, the dissatisfactions of $x \in C_1 \cap C_2$ will not be the same in C_1 and in C_2 because they are computed using all elements of these sets. So, instead of providing a representation result, we formulate this implication as a proposition.

Proposition 2 *R5 implies that for any C_1 and C_2 with $u[C_1^*] = u[C_2^*]$ it is true that $F_i(x \mid C_1) = F_i(x \mid C_2)$ for all $x \in C_1 \cap C_2$.*

Proof By the property of u^*.

Next, we aggregate the dissatisfactions across players and define F. This aggregation procedure combines individual normative judgments to generate a shared normative agreement that assigns relative appropriateness to each outcome according to the moral rule. We start by assuming that $F(x \mid C)$ is a function of $F_i(x \mid C)$ for all $i \in N$. Specifically, we assume that $F(x \mid C) = E(F_1(x \mid C), ..., F_N(x \mid C))$, where $E : \mathbb{R}^N \to \mathbb{R}_+$ is increasing in all arguments. The following axioms determine how aggregation proceeds.

R6 $E(0, ..., 0) = 0$.

R6 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 (R7) 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 $\omega_i \in (0, 1]$. 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 (see Kimbrough and Vostroknutov, 2021, for modeling details).[8]

R7 $\forall i \in N \, \forall t_1, ..., t_N \in \mathbb{R}_+ \, \forall a_i \geq -t_i \quad E(t_i + a_i; t_{-i}) = E(t_i; t_{-i}) + \omega_i a_i$.

The notation $E(t_i; t_{-i})$ singles out the ith argument of E. R7 says that if player i's personal dissatisfaction changes by a_i, then the aggregate dissatisfaction changes by

[8] There we also briefly discuss the implication that real-world normative disagreements can often be fruitfully understood in terms of disagreements about these weights. See (Kimbrough and Vostroknutov, 2022a) for a discussion of where these weights might come from.

the same amount, weighted by ω_i. R7 incorporates the idea that norms are more sensitive to changing dissatisfactions of "important" players with high ω_i, as compared to "unimportant" ones with low ω_i. The following proposition puts all the axioms together.

Proposition 3 The following two statements are equivalent:

1. f_i satisfies R1-R4, F_i satisfies R5, F satisfies R6-R7.
2. F can be expressed as

$$F(x \mid C) = \sum_{i=1}^{N} \omega_i F_i(x \mid C) = \sum_{i=1}^{N} \omega_i (\max\{u_i^*(x) - u_i(x), 0\} + \beta_i \max\{u_i(x) - u_i^*(x), 0\}),$$

where $u_i^*(x)$ *is short for* $u_i^*(u(x) \mid u[C^*])$ *and* $\beta_i \in (0, 1]$ *are coefficients.*

Proof See Appendix A.

The representation in Proposition 3 provides a simple mathematical expression for computing dissatisfaction for any allocation x in any context C. The norm-dependent utility of player i associated with $F(x \mid C)$ can be then defined as

$$w_i(x \mid C) = u_i(x) + \phi_i[-F(x \mid C)],$$

where $[\cdot]$ stands for the linear normalization of the function within to the interval $[-1, 1]$.[9] The main point is simple: player i trades-off personal consumption utility $u_i(x)$ and a linear function $[-F(x \mid C)]$ of the *negative* of aggregate dissatisfactions (since they ought to be minimized). They make this tradeoff according to their idiosyncratic propensity to follow norms, ϕ_i. In Kimbrough and Vostroknutov (2021) we call $\eta(x \mid C) = [-F(x \mid C)]$ a normative valence of x in C and η a norm function. Thus we can straightforwardly connect our model of moral reasoning to choice. An important implication of this model is that, in disinterested settings, where a decision-maker's payoff is entirely independent of their choices, normative considerations will be the sole determinant of those choices. This renders 3rd party allocation tasks especially useful for studying norm-driven behavior, an implication we take advantage of in Section 3.2 below.

[9] In Kimbrough and Vostroknutov (2021) we provide more details about why we need this normalization.

3.1 Constructing Aggregate Dissatisfaction for Moral Rules

The following examples show how to express some commonly studied moral rules using a dissatisfaction function F that satisfies the axioms above. In each case, we do not exactly specify F but rather its inputs: C^* for each C and the function u^*. For simplicity, we assume in all examples that $\omega_i = 1$ for all $i \in N$.

Pareto Optimality For all C we have $C^* \subseteq C$, and for all $x, y \in C$, if $u(x) \geq u(y)$, with strict inequality for at least one component, then $y \notin C^*$. Thus, C^* is non-empty for all C, and $u^*(r \mid S)$ can be defined as an element of S that is the closest to r in Euclidean metric.

Cooperative Solution Concepts Many cooperative solution concepts can be expressed as moral rules. To illustrate the idea, we take the general class of bargaining solutions in Karos et al. (2018) that includes the Nash Bargaining solution and the Kalai-Smorodinsky solution. These solutions, characterized by $0 \leq p < \infty$ and defined on convex sets S, pick an allocation that is weakly Pareto optimal and have the ratio of payoffs for any two players i and j equal to the ratio of their maximal payoffs in S raised to the power p and denoted by $a_i^p(S)/a_j^p(S)$, where $a_i(S)$ is the maximal payoff that i can get in S. We can express this in our framework by defining for each set C the ideal singleton set $C^* = s^*(C)$, where $s^*(C)$ is the point on the ray going through the origin and the point $(a_1^p(C), ..., a_i^p(C))$, such that $s_i^*(C) = a_i(C)$ for some $i \in N$. This way, for $p = 0$ we capture a moral rule based on the Nash Bargaining solution and for $p = 1$ we get a moral rule based on the Kalai-Smorodinsky solution.

Payoff Efficiency For all C we have $C^* \subseteq C$, and for all $x, y \in C$, if $\sum_{i \in N} u_i(x) > \sum_{i \in N} u_i(y)$, then $y \notin C^*$. $u^*(r \mid S)$ can be defined as an element of S that is the closest to r in Euclidean metric.

Maximin For all C we have $C^* \subseteq C$, and for all $x, y \in C$, if $\min_{i \in N} u_i(x) > \min_{i \in N} u_i(y)$, then $y \notin C^*$. $u^*(r \mid S)$ can be defined as an element of S that is the closest to r in Euclidean metric.

Choice-Set-Dependent Inequality Aversion Take any C and compute the average payoff that each player gets in all consequences: $a_i^* = \sum_{c \in C} u_i(c)/|C|$. Let $C^* =$

$\{(a_i^*)_{i\in N}\}$. This is a singleton ideal set, which is sensitive to all payoffs that players can receive. In this case, there is no need to define u^*.[10]

It is also worth considering the simplest class of moral rules for which the ideal set C^* is independent of C. In this case, for any non-ideal $x \in \mathcal{C}$, the dissatisfaction from x is the same for all $C \in \mathcal{C}$ that include x. In other words, $F_i(x \mid C) = F_i(x) = f_i(u_i(x), u^*(u(x) \mid u[C^*]))$. This moral rule closely approximates outcome-based social preference models, with the only difference being that in our approach all players attach the same aggregate dissatisfaction $F(x)$ to an outcome, whereas in social preference models different players can have different social utility terms at the same outcome (e.g., inequality aversion à la Fehr and Schmidt, 1999). Our model instead incorporates heterogeneity through the parameter ϕ_i in the norm-dependent utility function.

The following example illustrates a choice-set-independent dissatisfaction function.

Choice-Set-Independent Inequality Aversion Let $C^* = \{x \in \mathcal{C} \mid \forall i, j \in N \; u_i(x) = u_j(x)\}$. This is a ray in $\mathbb{R}^N$ containing all allocations that give the same payoff to all players. For any $x \in \mathcal{C}$ and $i \in N$, let $a = \sum_{j\in N} u_j(x)/N$ and $u_i^*(u(x) \mid u[C^*]) = (a, ..., a) \in C^*$ be the average payoff (identical for all players).

3.2 An Example

To build intuition, we illustrate the strengths and weaknesses of modeling moral reasoning via moral rules with two simple examples. Consider two Dictator games discussed in List (2007). In one, dictators choose an offer $x \in [0, 5]$ that defines an allocation $(5 - x, x)$, where $5 - x$ goes to the dictator. This represents the standard Dictator game (only giving options). In the other, the dictator still chooses an allocation $(5 - x, x)$, but now the dictator's payoff is $x \in [-5, 5]$, meaning the dictator may both give and take. List (2007) finds that in the former game, subjects frequently give 2.5 (the midpoint between 0 and 5) often achieving the equal allocation $(2.5, 2.5)$. By contrast, in the latter game subjects are much more likely to choose 0 (the midpoint between -5 and 5), generating a very unequal allocation $(5, 0)$. Out of all the moral rules described in the previous section, only choice-set-dependent inequality

[10] A similar but less payoff-sensitive principle of inequality aversion can be specified if we take into account only the highest and the lowest payoffs that players receive in C. For each player compute $\underline{m}_i^* = \min_{c\in C} u_i(c)$ and $\overline{m}_i^* = \max_{c\in C} u_i(c)$. Let $C^* = \{(\frac{\underline{m}_i^* + \overline{m}_i^*}{2})_{i\in N}\}$. This is again a singleton ideal set. However, now it is less choice-set-dependent: adding any consequences that do not change $\underline{m}_i^*$ and $\overline{m}_i^*$ does not change the ideal element and thus does not affect the dissatisfaction distance of existing elements in C.

aversion accounts for this behavior. Other concepts either fail to predict treatment effects (Pareto optimality, payoff efficiency), or predict a different effect (maximin, cooperative concepts, choice-set-independent inequality aversion).

From this example, it may seem that choice-set-dependent inequality aversion is the moral rule that fits these data the best. Therefore, it seems reasonable to assume that people are more likely to use this specific rule than others. However, the example in Figure 1 shows that choice-set-dependent inequality aversion is not always a reasonable moral rule. The left panel of the figure represents a Dictator game with a pie size of 3 and available allocations A, B, D and E. According to choice-set-dependent inequality aversion, the allocation C^* is the ideal for this context. This implies that the allocations B and D (marked with white circles) are the most appropriate since they are the closest to C^*.

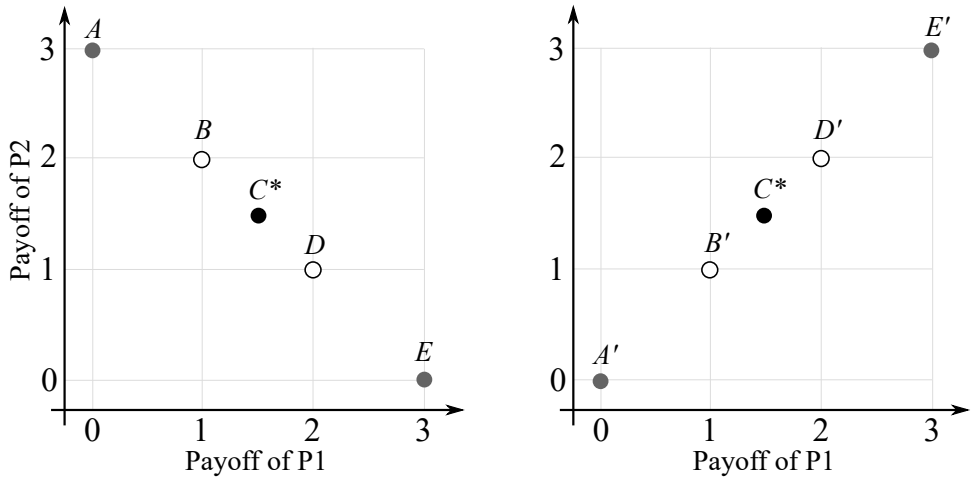

Figure 1: *Left.* Choice-set-dependent inequality aversion in the context of the Dictator game with allocations A, B, D, E (the ideal is C^*). White circles denote the most appropriate allocations with respect to C^*. *Right.* Same but in the context A', B', D', E' obtained from the Dictator game by redefining allocations (with the same C^*).

Now consider another context shown in the right panel of Figure 1. This context consists of allocations A', B', D' and E' that were obtained from the allocations A, B, D and E by taking both players' payoffs in these allocations, ranking them, and creating new allocations from the ranked payoffs.[11] We call this game the Efficiency Game (EG) since payoffs at each outcome are equal across players but the outcomes differ in their efficiency. Note that—according to the moral rule that we call choice-set dependent inequality aversion—C^* is the same in the EG as it was in the Dictator

[11] This is a general procedure that can be applied to any context. For some context C, let $a_1^i \geq a_2^i \geq ... \geq a_{|C|}^i$ be the ranked payoffs of player i in C. Then the new context is obtained from the old one by defining allocations $(a_k^1, ..., a_k^N)$ for $k = 1..|C|$.

game. According to the moral rule, C^* is computed by averaging the payoffs over all outcomes for each individual separately, and in both contexts the same set of payoffs are being averaged. While B and D are intuitively appealing in the DG, the fact that C^* remains the same in the EG obviously clashes with normative intuition. It seems unlikely that people would choose B' or D', which are the most appropriate allocations from the perspective of choice-set-dependent inequality aversion (marked with white circles).

Thus, we conducted an online experiment on Prolific with 100 people from the UK testing this intuition. Subjects made 3rd party allocation decisions in the DG and EG (in random order) where their choice influenced others' payoffs but had no influence on their own payoff. As noted above, under norm-dependent utility, a 3rd party allocation decision should be driven entirely by normative considerations, since the decision-maker is disinterested. Thus we also used the coordination-game method due to Krupka and Weber (2013) to elicit shared beliefs about the appropriateness of each action in each game. Subjects were incentivized to guess the most common response given by others; if a commonly known injunctive norm exists, then subjects can resolve the coordination problem by using that norm as a focal point.[12]

The results are shown in Figure 2. The gray bars show the proportion of subjects choosing each allocation, and the black lines show the average appropriateness of each allocation (measured on a 4-point Likert scale from Very Socially Inappropriate to Very Socially Appropriate). Consistent with choice-set dependent inequality aversion, the more egalitarian options, (1,2) and (2,1) are rated as the most appropriate actions in the 3rd party DG; moreover, choices are approximately equally split between these options. In the 3rd party EG, the results are different; subjects clearly favor the more efficient outcome, both normatively and as reflected in their choices. Thus, choice set dependent inequality aversion, though capable of accounting for DG choices, is unable to account for EG choices, as efficiency considerations take over.

This example highlights an important limitation of theories that codify a particular moral rule and assume that it applies to all choice contexts. Although such moral rules exhibit a degree of context-dependence, they still commit decision-makers to apply a single notion of the ideal. This will render the norms derived from such a model unable to account for instances in which people seem to reason using different

[12]We chose one of the 4 tasks at random to count for payment; subjects received 15 pence for participating in the 2-minute study plus their earnings from the randomly chosen task. Subjects were told that, if the coordination game was the task for which they were paid, we would select one action from the EG or DG at random. Then, if their response matched the modal response for that action, they received 1GBP; otherwise they received nothing. The Qualtrics questionnaire is available in Appendix C. This experiment was conducted with approval from the Maastricht intercity ethics committee under their common IRB agreement with the BEELab at Maastricht University ($ERCIC\_379\_28\_09\_2022$).

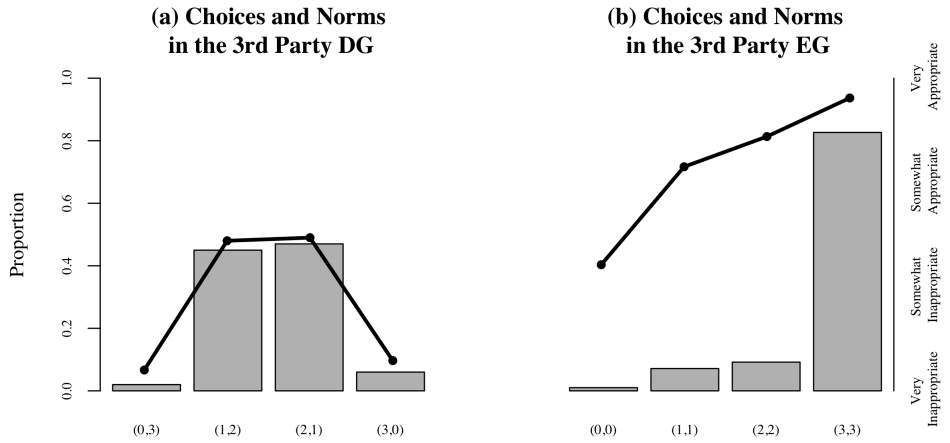

Figure 2: Choices and Elicited Norms in the 3rd Party DG and EG Shown in Figure 1

normative principles in different contexts. The example suggests that a more *radical context dependence* is necessary to account for the observed diversity of social behavior. We need a model rooted in moral psychology, in which the moral rule itself arises from the choice context. In the next section, we propose one such model.

4 Endogenizing the Moral Rule

Here, we provide an alternative set of axioms rooted in moral psychology that endogenizes the moral rule to the choice context. We build on some of the assumptions made in the R-model, but we replace key axioms with alternatives that allow moral judgment to depend on context in a rather radical, but intuitively plausible, way. In particular, 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.[13]

Thus when defining dissatisfaction, we assume that all agents are dissatisfied when they are unable to achieve outcomes that are selfishly better for them, rather than specifying an abstract ideal to which other outcomes are compared. 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." Our axioms are inspired by an account of moral psychology drawn from Smith (1759) and elaborated in Smith and Wilson (2019)

[13]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 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.

in which our account of what is right/good/appropriate emerges from our experience with what others approve and disapprove of, rather than the other way around.

Next, we define the set of *P-axioms* (for moral psychology) that incorporate this psychological intuition and generate the *P-model*. The notation is similar to the R-model but with d_i denoting dissatisfaction functions instead of f_i; personal dissatisfactions becoming D_i instead of F_i; and aggregate dissatisfaction becoming D instead of F.

P1 $\forall i \in N \; \forall t, r, a \in \mathbb{R} \quad d_i(t + a, r + a) = d_i(t, r)$.

P1 is exactly the same as R1, stating that adding a constant to the utilities being compared does not change the dissatisfaction. So, only the difference in utilities matters.

P2 $\forall i \in N \; \forall t, r \in \mathbb{R}$ with $t \geq r$ we have $d_i(t, r) = 0$.

P2 differs from R2 in that it introduces an asymmetry between prospective gains and prospective losses. In particular, P2 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 normative valence of the superior consequence.[14] 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 dissatisfaction derived from superior outcomes.

P3 $\forall i \in N \; \forall t, r, \alpha \in \mathbb{R}$ with $\alpha > 0 \quad d_i(\alpha t, \alpha r) = \alpha d_i(t, r)$.

This axiom is the same as R3. Finally, as above we assume non-triviality and equivalence of dissatisfactions across players:

P4 $\forall i \in N \quad d_i(0, 1) = 1$.

The following proposition establishes the functional form of d_i equivalent to axioms P1-P4.

Proposition 4 The following two statements are equivalent:

1. d_i satisfies P1-P4;
2. $d_i(t, r) = \max\{r - t, 0\}$.

[14] 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 A.

Next, we introduce axioms that define the personal dissatisfaction D_i of player i associated with a single consequence x.

P5 $\forall i \in N \, \forall x \in \mathcal{C} \quad D_i(x \,|\, \{x\}) = 0.$

Axiom P5 states that if only one consequence is available or possible, then there is nothing to be dissatisfied about, so the dissatisfaction of each player i is zero. Although P5 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 $\{x\}$.[15]

P6 $\forall i \in N \, \forall C \subset \mathcal{C}, x \in C, y \in \mathcal{C} \backslash C$

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

Axiom P6 defines the personal dissatisfaction function of player i. 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. We think of P6 as capturing another basic sentiment, namely that player i feels dissatisfaction whenever a new possibility y appears, that could give i a higher payoff.

The following result connects the axioms above and the representation that we test in Kimbrough and Vostroknutov (2021).

Proposition 5 The following two statements are equivalent:

1. D_i satisfies P5-P6;

[15] 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 P5 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, 2021). 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.

2. D_i can be expressed as $D_i(x \mid C) = \sum_{c \in C \setminus \{x\}} d_i(u_i(x), u_i(c))$.

Proof See Appendix A.

Finally, we define aggregated dissatisfaction $D(x \mid C)$. This is done in the same way as in the R-model. We assume that D is a function of D_i's: $D(x \mid C) = G(D_1(x \mid C), ..., D_N(x \mid C))$. The following two axioms describe the properties of G.

P7 $G(0, ..., 0) = 0$.

This is the same as R6. The last axiom is the same as R8 assuming the existence of social weights $(\omega_i)_{i \in N}$.

P8 $\forall i \in N \; \forall t_1, ..., t_N \in \mathbb{R}_+ \; \forall a_i \geq -t_i \quad G(t_i + a_i; t_{-i}) = G(t_i; t_{-i}) + \omega_i a_i$.

The proposition below puts all the axioms together.

Proposition 6 The following two statements are equivalent:

1. d_i satisfies P1-P4, D_i satisfies P5-P6, D satisfies P7-P8.
2. D can be expressed as

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

Proof Similar to the proof of Proposition 3.[16,17]

[16] The representation in Proposition 6 contains the maximum of a utility difference and zero, which might remind many readers of the inequality-averse utility function introduced by Fehr and Schmidt (1999). However, our model is conceptually different: inequality aversion considers utility differences at a given outcome *across players*; our model of dissatisfaction considers utility differences of the same player *across outcomes*. Then we aggregate these across players to define the injunctive norm.

[17] In the case with only one player ($N = 1$) and keeping in mind the result of Proposition 4, 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 the P-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 utility in choice.

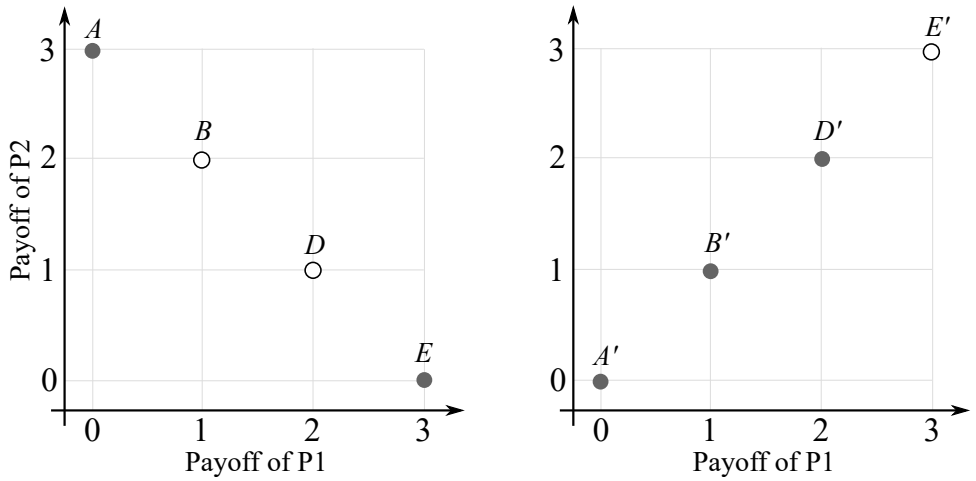

Figure 3: *Left.* P-model in the Dictator game. White circles denote the most appropriate allocations. *Right.* Same but in the context A', B', D', E' obtained from the Dictator game by redefining allocations.

Next, to illustrate the contrast to the R-model we show how the P-model evaluates outcomes in the same example we used in the Section 3. Figure 3 shows the same allocations, and the white circles mark the most appropriate allocations from the perspective of the P-model.

According to the P-model, in the left panel of Figure 3 the allocations B and D are the most appropriate because these allocations balance the interests of P1 and P2 (see Proposition 10 below for the proof). In the right panel, the allocation E' becomes the most appropriate (it Pareto-dominates all others, see Proposition 8 below). While the rankings in the left panel are consistent with choice-set-dependent inequality aversion, the ranking of outcomes in the right panel are not. However, the P-model rankings *are* consistent with the ones supplied by subjects in the experiment reported in Figure 2. In our companion paper, we show that the radical context dependence implied by P-axioms can also explain the shifts in normative perspective documented in a variety of previous experiments (see Kimbrough and Vostroknutov, 2021, for more details).

5 Non-Equivalence of the P- and R-Models

In this section we show that norms generated from moral rules in any R-model are less context-dependent than norms generated by the P-model. To do this we need a notion of equivalence between them. Given that the d and f functions in P- and R-models are not fixed, it might seem as though we could focus on whether they have the same maximal element. However, the entire ranking of outcomes is relevant for choice because interior solutions of the utility maximization problem with norm-dependent prefer-

ences are sensitive to the relative normative valences of all the outcomes. Thus, we compare the models by asking whether they induce the same ranking of consequences in terms of dissatisfaction.

Let $\succcurlyeq_C$ be defined as the preference relation that represents the aggregated dissatisfactions in the P-model in some set C:

$$\forall C \in \mathcal{C} \quad x \succcurlyeq_C y \Leftrightarrow D(x \,|\, C) \geq D(y \,|\, C).$$

Similarly, for some R-model, defined by the collection of all C and C^* together with a distance function f, let $\succcurlyeq_C$ represent the dissatisfactions:

$$\forall C \in \mathcal{C} \quad x \succcurlyeq_C y \Leftrightarrow F(x \,|\, C) \geq F(y \,|\, C).$$

Let us say that a R-model is *equivalent* to the P-model if $\succcurlyeq_C$ and $\succcurlyeq_C$ are the same for all $C \in \mathcal{C}$. Our task now is to show that there exists no R-model that is (1) equivalent to the P-model and (2) satisfies all P-axioms. Let us try to construct a R-model that is as close as possible to the P-model. R-models are defined through ideal elements C^* and a distance function. So, for each C let us define C^* as the set of minimal elements of $\succcurlyeq_C$ for each C (the P-model preference relation) and choose any dissatisfaction function f.

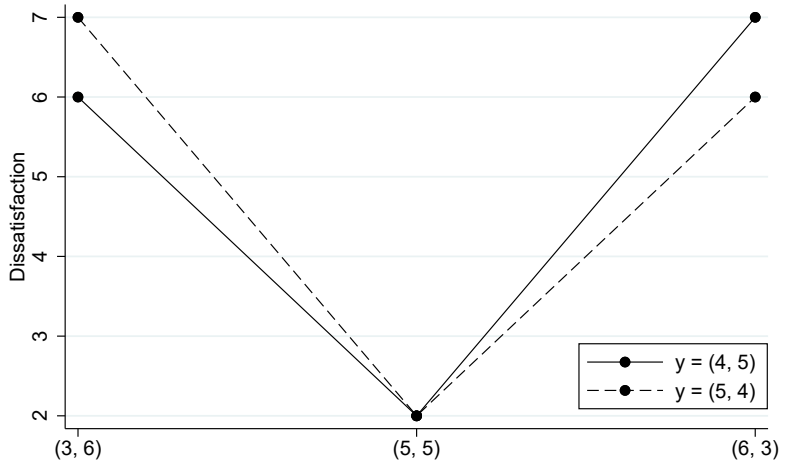

Figure 4: P-model dissatisfactions of the common allocations in $C_y = \{(3,6);(5,5);(6,3);y\}$, where $y = (4,5)$ or $y = (5,4)$.

The following example shows how no such model can be equivalent to the P-model. Consider two sets of four allocations for two players $C_y = \{(3,6);(5,5);(6,3);y\}$ where in one set $y = (4,5)$ and in the other $y = (5,4)$. In the P-model both $C_{(4,5)}$ and $C_{(5,4)}$ have minimal dissatisfaction at $(5,5)$. So, when we construct a R-model as described above for both of them we set $C^* = \{(5,5)\}$. By Propositions 2 and 3, in any

R-model $F((3,6) \mid C_{(4,5)}) = F((3,6) \mid C_{(5,4)})$ and the same holds for allocation $(6,3)$. Thus, according to any R-model either $(3,6)$ or $(6,3)$ has *larger* dissatisfaction in *both* $C_{(4,5)}$ *and* $C_{(5,4)}$. However, Figure 4 shows that this is inconsistent with the P-model in which the relative dissatisfactions of $(3,6)$ and $(6,3)$ change depending on y. The dissatisfaction of $(3,6)$ is smaller than that of $(6,3)$ in $C_{(4,5)}$, but larger in $C_{(5,4)}$. Thus, no R-model is equivalent to the P-model. We state this result as a proposition.[18]

Proposition 7 *There is no R-model that is equivalent to the P-model in the sense of generating equivalent preference orderings in terms of dissatisfaction for all contexts.*

Proof By example on Figure 4.

This proposition demonstrates that if aggregate dissatisfaction is what people care about—as the P-model postulates—then there is no R-model that reflects the same criterion. We have shown that this is the case with the class of R-models that are the closest to the P-model, namely those that have the same minimal dissatisfaction as the P-model in all sets C.

For more general R-models the discrepancy is even larger. If $C^* \subsetneq C$, or there are elements of C^* outside of C, then adding these elements to C does not change any dissatisfactions. However, in the P-model, in general, the dissatisfactions of all elements will change after such an addition, and the minimal dissatisfaction will not be assigned in the same way by the P- and R-models.

Moreover, the dissatisfaction-minimizing allocations in the P-model are never Pareto-dominated (see Proposition 8 below). Thus, if in a R-model some C^* contains Pareto-dominated consequences, as for example in the choice-set-dependent inequality aversion, then this will always contradict the predictions of the P-model except in special cases.

Finally, the discrepancies in induced preference orderings between P- and R-models are not "rare" or measure zero in some sense: a reversal in the dissatisfaction rankings of some allocations (as demonstrated in Figure 4) can be very easily constructed and is rather typical for the kind of context-dependence that arises from the P-model. Thus, we should expect systematic differences between P- and R-models in terms of dissatisfaction rankings.

[18] It may seem that we could construct an equivalent R-model by creating an ideal consequence that gives each player the highest payoff that they can receive at any element in C, but because the normative evaluation of each outcome depends on all other outcomes (and not just on the ideal), this will not generate an equivalent ranking.

6 Some Properties of P-Model

In Kimbrough and Vostroknutov (2021), we test a variety of implications of the P-model using existing experimental data and show that it provides a powerful framework to explain context-dependent social behavior. Here, we provide some theoretical results on the properties of the norms derived from the P-model. This serves to establish some regularities, highlighting that the P-model has structure despite its radical context dependence, and provides intuition for the kinds of context dependence that emerge naturally from it.[19]

The first result states that the injunctive norms implied by the P-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 P-model always satisfy the Pareto optimality criterion. We formulate this as a proposition.

Proposition 8 *In any context $C \subset \mathcal{C}$, if $c_1 \in C$ Pareto dominates $c_2 \in C$ then $D(c_1|C) < D(c_2|C)$.*

Proof See Appendix A.

It is worth noting that our model can also provide a ranking across Pareto optimal elements of the choice set. Thus, although all dissatisfaction-minimizing consequences are Pareto optimal, the converse is not true, and the ability of the Pareto optimality criterion to predict the normatively best consequence (under the P-model, for example) is limited.

In fact, we prove the following relatively general result that shows that the normatively best consequences under the P-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, can be seen as a kind of resource constraint. We provide formal definitions and proofs of this result in Appendix B. Here we simply offer a sketch of the proposition to build intuition.

Proposition 9 (Midpoint Theorem) *Suppose that C is an N-dimensional convex*

[19] In a followup paper, Kimbrough and Vostroknutov (2022b) computationally evaluate the ranking of outcomes induced by the P-model in a variety of decision making contexts and ask whether, for particular kinds of choices, the P-model can be reasonably summarized by particular moral rules. In other words, we attempt to assess to what extent the P-model can be thought of as a meta-theory of moral rules, telling us which moral rules people may be likely to articulate in which contexts.

polytope in $\mathbb{R}^N$ that satisfies the Scarcity Condition. Then, the normatively best consequence according to the P-model is such that no one player gets the maximal possible payoff.

Proof See Appendix B.

Proposition 9 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 P-model generally predicts that all players need to compromise and sacrifice something for the sake of achieving a cooperative 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 Midpoint Theorem for contexts involving only two players and constant efficiency, but which nonetheless provides a more specific characterization of the most appropriate allocation.

Proposition 10 *Suppose there are two players and K consequences $C = \{c_1, c_2, ..., c_K\}$ with utilities $u_1 \leq u_2 \leq ... \leq u_K$ for one player and $a - u_1 \geq a - u_2 \geq ... \geq a - u_K$ for the other ($a, u_1, ..., u_K \in \mathbb{R}$). Then, for any $j = 1..K-1$, $D(c_{j+1}|C) - D(c_j|C) = (2j - K)(u_{j+1} - u_j)$. Thus, the midpoint consequences $c_{\frac{K}{2}}$ and $c_{\frac{K}{2}+1}$, if K is even, and $c_{\frac{K}{2}+\frac{1}{2}}$, if K is odd, have the smallest dissatisfaction in the P-model.*

Proof See Appendix A.

Proposition 10 implies that the most appropriate consequence under the P-model, in cases with constant payoff efficiency of all possible allocations, is not the one that is the closest to an equal distribution of utility, as most models of social preferences would suggest, but rather the one that is "equal" in terms of the number of other undesirable consequences available: for the most appropriate consequence this number is the same for both players. Thus, consequences that yield very unequal utilities across agents, can still be considered normatively appropriate in specific contexts where most consequences give a large portion of the pie to one player.

6.1 P-Model in Social Dilemmas

Next, we establish some theoretical results about the P-model for commonly studied games in which normative motivations arguably play an important role in decision-

making. 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 the P-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 (see Example 6 in Kimbrough and Vostroknutov (2021) for full analysis).

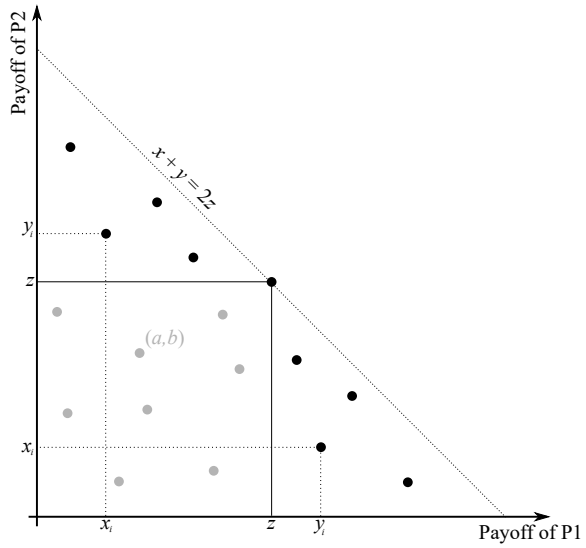

Figure 5: Symmetric two-player social dilemma.

Here, we provide formal analysis that generalizes this point, considering a class of symmetric two-player social dilemmas in which the P-norm corresponds to the most efficient outcome—even though for both players, there exist other consequences that bring them higher utility. Consider the set of payoff allocations for two players shown in Figure 5. 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 under the P-model.

Proposition 11 *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 in the P-model.*

Proof See Appendix A.

To provide additional intuition about the kind of behavior implied by the injunctive norm in social dilemma games, consider Figure 6 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.

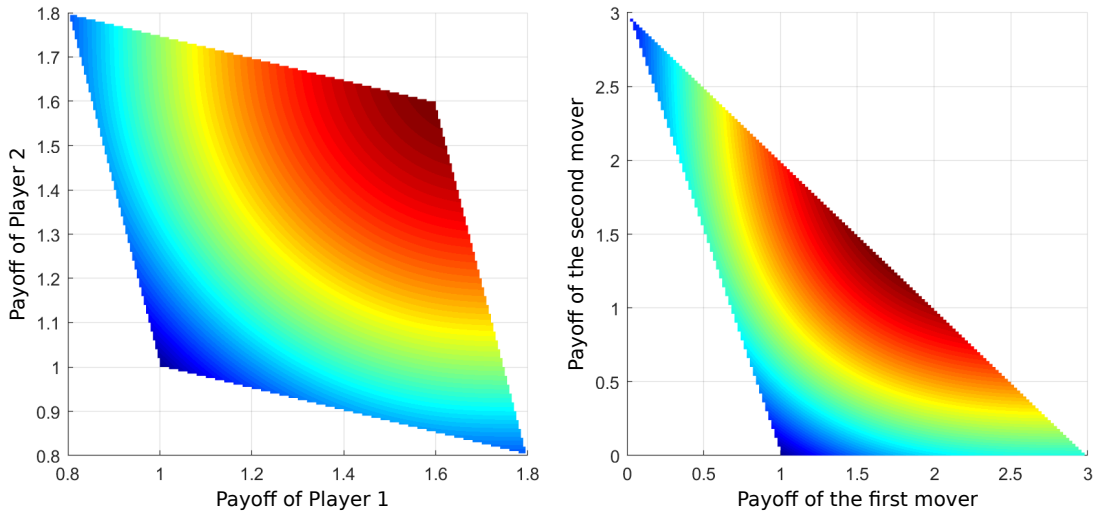

Figure 6: *Left.* The injunctive norm derived from the P-model in a 2-player Public Goods Game. *Right.* The injunctive norm derived from the P-model in a Trust Game. Dark red shows the most appropriate consequence and deep blue shows the least appropriate one.

The most appropriate consequence in the Public Goods game is for both players to contribute the whole endowment, which follows from Proposition 11, and the most inappropriate consequence is to contribute nothing. The normatively best outcome maximizes both efficiency and equality in this case.[20]

[20]Technically, Proposition 11 holds for finite sets of allocations and does not apply to the allocations

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 11. It is also worth noting that for any choice of the first mover, the P-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. We discuss the connection to reciprocity more thoroughly in Kimbrough and Vostroknutov (2021).

6.2 Positive Reciprocity In Extensive-Form Games

In this section we demonstrate how positive reciprocity can emerge from P-model in extensive-form games. By positive reciprocity we mean the tendency to return a favor. We do not consider negative reciprocity here. In our framework it is incorporated as punishment of norm violators (Kimbrough and Vostroknutov, 2021).

We begin by discussing some conceptual results from Isoni and Sugden (2018). In this paper the authors (IS) analyze the simple two-player, two-move game shown in Figure 7. IS consider an ideal Trust World in which Player 1 chooses *send* and Player 2 chooses *return*, both with probability 1 (the game on the left), while at the same time Player 2 chooses *equal* with probability less than 1 in the restriction of this game that does not include the move of Player 1 (the game on the right). According to IS, the idea of trust and trustworthiness is that in the game on the left, Player 2 chooses *return* with higher probability than she chooses *equal* in the game on the right, *exactly because* Player 2 enters a trust relationship with Player 1 when he chooses *send*.

IS note that most models of reciprocal kindness do not support the aforementioned strategies as an equilibrium. They call this the Paradox of Trust. These models cannot account for trust in this basic game because of the way they model reciprocity: players are assumed to respond with kindness by others with kindness of their own. The problem is that the action *send* cannot be classified as either kind or unkind, since Player 1 chooses it *expecting* that Player 2 chooses *return*. Thus both a kind Player 1 and a selfish Player 1 who wants to maximize utility will do the same thing. IS argue that this contradicts the idea of kindness, which holds that an action is kind only if it is done *without* expectation of something in return. IS conclude that trust behavior in

sets in Figure 6 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 11 for continua can be formulated. We leave this for the future research.

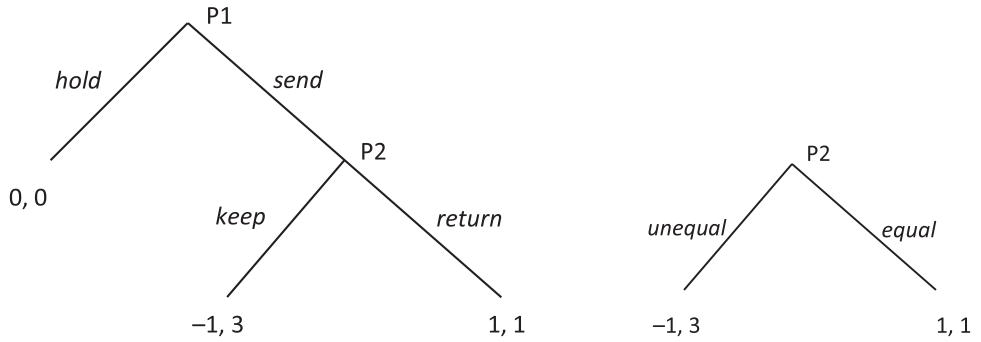

Figure 7: *Left.* The Trust game considered in Isoni and Sugden (2018). *Right.* The Dictator game faced by the second player in the absence of the move of the first player.

this game cannot be based on reciprocal kindness, which presumes a reaction by Player 2 to the known intentions of Player 1, because Player 1's intentions cannot be inferred from his actions. They propose instead that this type of trust should instead be thought of as a "joint action" of the players who are involved in "reciprocal cooperation" (Isoni and Sugden, 2018).

Our P-model works as IS suggest. The consequence $(1, 1)$ in the game on the left is the most appropriate, so any players who care enough about following norms choose *send* and *return* in a "joint enterprise," which is to behave in the socially appropriate way according to the P-model. The norm is shared, so in the limit as their propensity to follow norms goes to infinity, they are aiming at exactly the same thing. In the Dictator game on the right, the two actions of Player 2 have the same normative valences, so according to norm-dependent utility, she chooses the selfish option *unequal*, again exactly as IS hypothesize. Our theory thus resolves the Paradox of Trust inherent in the models of reciprocal kindness and shows that trust can be based on social norms, which *create* reciprocal behavior exactly because norms are the common factor (joint enterprise) that enters players' norm-dependent utilities.

We support the idea that reciprocal behavior arises more generally from the P-model with an example and a proposition. 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. Notice as well that the moves of the two players can be separated in time, incorporating the idea of a "contract" into which the players enter (Isoni and Sugden, 2018).

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 $(p_1^*, p_2^*) = \arg\max_{(p_{1i}, p_{2i})} p_{1i} + p_{2i}$. In the following proposition we show that choosing a^* by both players leads to the outcome which is the P-norm, or to the outcome with the smallest P-dissatisfaction.

Proposition 12 *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 P-dissatisfaction.*

Proof See Appendix A.

Proposition 12 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 P-norms create a pattern of reciprocation with the same action a^*. This idea 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 P-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 the P-model.[21]

6.3 P-Model in Coordination Games

In coordination games, the normatively best outcome under the P-model depends on what other assumptions we make about payoffs. In coordination games with multiple

[21] We would like to mention again at this point that negative reciprocity in our framework is conceptually distinct. Specifically, in Kimbrough and Vostroknutov (2021) we construct a mechanism for punishment of norm violators that incorporates an *Eye for an Eye* principle, which means that the magnitude of punishment is proportional to the extent of a norm violation. This, in our view, represents negative reciprocity, which is driven by the desire to punish rather than "reciprocate" in the sense of the behavior in the gift exchange game. "Negative reciprocity" in a gift exchange game emerges when the first mover does not choose the action compatible with reaching the normatively best outcome and gets punished for this by the second mover.

Pareto-ranked equilibria (minimum effort games, stag hunts), the P-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 P-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' utilities 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). In Kimbrough and Vostroknutov (2022a) we discuss in more detail how social weights can come about in social interactions.

6.4 Constant-Norm Environments

While the P-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 P-model is constant across all feasible consequences. Let the tuple $\langle N, C, u\rangle$ that consists of the set of players, some set of consequences C, and a utility function u be called an *environment*. Call an environment $\langle N, C, u\rangle$ *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 $\langle N, C, u\rangle$ 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 13 *Any tournament is constant-norm.*

Proof See Appendix A.

Thus, built into our P-model, which accounts for the dissatisfaction of all interested parties, is the existence of situations in which moral reasoning does not single out any consequence as more appropriate than others. Under our model of norm-dependent utility, players in such settings are expected to behave as self-interested utility maximizers, which seems like a reasonable prediction of players' motivation in tournaments. 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.

7 Discussion

Our paper is motivated by the mounting evidence that economic decision-making reflects not only self-interested payoff maximization but also normative considerations that allow us to cooperate, coordinate, and pursue common goals. This evidence has led to fruitful modeling efforts that capture these often-conflicting motivations in a norm-dependent utility function, which models people as trading off what they *ought* to do against what they *want* to do from the point of view of consumption utility. This calls for formal theories of how agents determine what they *ought* to do in any situation. Our work addresses this call by bridging a gap between cooperative and non-cooperative game theory.

Our innovation is to reinterpret axiomatic models of the kind developed in cooperative game theory as theories of moral reasoning and norm formation. Typically such reasoning is conducted by modelers who use the normative rankings of outcomes from such models to assess whether a normatively desirable outcome (as derived from the axioms) has been achieved by agents who do not necessarily share this ideal. We argue that instead, it makes sense to put the moral reasoning into the minds of the

decision-makers themselves. This approach allows us to combine the best elements of cooperative and non-cooperative game theory, using the former to identify the kinds of allocations that norm-following agents will aim at and the latter to identify conditions under which those allocations are likely to actually be attained.

We illustrate the potential of this method by developing axiomatic models of moral reasoning rooted in the idea that people are dissatisfied when feasible outcomes would leave them with something other than what is normatively ideal. We first introduce axioms in which dissatisfaction is computed relative to some ideal specified by a moral rule: for example, maximize efficiency or minimize inequality. We show how to define such ideals in a way that closely approximates social preference models, and we show that, despite the intuitive appeal of moral reasoning based on a simple, general rule, such models are unable to account for some evidence of context dependence in choice.

Thus we propose an alternative set of axioms that constructs injunctive norms endogenously from the choice set, based on an intuitive theory of moral psychology. We highlight that injunctive norms derived from this model are radically choice-set-dependent. How appropriate a particular outcome is depends crucially on the other possible outcomes. In a companion paper (Kimbrough and Vostroknutov, 2021), we show that a functional form derived from this model has substantial explanatory power in experiments and in particular can account for many observations of radical context dependence in choice that cannot be readily explained by standard theory or in models of outcome-based social preferences.

One final distinction between moral rules in the R-model and the radically choice-set-dependent moral psychology of the P-model is worth drawing out. Choice-set dependence means that normative goals are always defined endogenously on the set of feasible allocations; moral rules, on the other hand, can define the ideal in terms of possibilities that exist outside the feasible set. Thus, moral rules can have an aspirational quality; when the ideal outcome is outside the feasible set, knowledge of a moral rule may induce people to seek changes to the feasible set.

Even if a reader emerges unpersuaded by our chosen axioms, we believe our framework has value in providing a new way to talk about the influence of normative considerations on decision-making. There are some very strong assumptions required to make the model work, such as complete knowledge of others' utility functions, shared axioms from which norms are derived, agreed-upon weights that are applied in the aggregation process, and so on. All of these are likely to be violated to varying degrees, in practice, but in our view, thinking about them in the language of our framework can help us understand the causes and consequences of normative disagreement more clearly.

For example, we can use the framework to make predictions about the kinds of settings in which people who assume different utility functions will see the same action in a normatively different light. Similarly, it allows us to engage in comparisons, identifying how different agents, motivated by different normative axioms should be expected to behave in various circumstances. On what will they agree? On what will they disagree? What is the potential scope of cooperation between agents with fundamentally different normative commitments? What will be the zones of conflict? Finally, our framework can help us understand instances of protest and conflict in terms of disagreements and negotiations over the weights that ought to be applied to various individuals and groups in our normative calculus. Marginalization might be defined in terms of the implied weight assigned to individuals or groups in social decisions. Demands for inclusion and representation can be understood as an attempt to assert the right of such marginalized people to be counted equally. We hope our framework can provide a language to talk about such issues and to incorporate agents who care about them, and other normative concerns, into economic modeling.

References

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

Alger, I. and Weibull, J. W. (2013). Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica*, 81(6):2269–2302.

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.

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.

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

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. (2018). Efficiency versus equality in bargaining. *Journal of European Economic Association*, forthcoming.

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

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

Karos, D., Muto, N., and Rachmilevitch, S. (2018). A generalization of the Egalitarian and the Kalai–Smorodinsky bargaining solutions. *International Journal of Game Theory*, 47(4):1169–1182.

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

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

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

Kimbrough, E. and Vostroknutov, A. (2021). A theory of injunctive norms. mimeo, Chapman University and Maastricht University.

Kimbrough, E. O. and Vostroknutov, A. (2022a). Affective decision-making and moral sentiments. mimeo, Chapman University and Maastricht University.

Kimbrough, E. O. and Vostroknutov, A. (2022b). Moral psychology as a meta-theory of moral rules. Working Paper.

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.

Nash, J. (1950). The bargaining problem. *Econometrica*, 18(2):155–162.

Nash, J. (1953). Two-person cooperative games. *Econometrica*, 21(1):128–140.

Perry, M. and Reny, P. J. (1994). A noncooperative view of coalition formation and the core. *Econometrica*, 62(4):795–817.

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

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

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

Appendix (for online publication)

A Proofs

Proof of Proposition 1 $(1 \Rightarrow 2)$. By R2, $f(0,0) = 0$ and by R1, $f_i(t,r) = f_i(0, r-t)$. So if $r - t \geq 0$ then $f_i(t,r) = f_i(0,1)(r-t) = r - t$. The last equality is by R4. When $r - t < 0$, by R2 and the above we have $f(t,r) = \beta_i f_i(0, t-r) = \beta_i(t-r)$. So, together we can write

$$f_i(t,r) = \max\{r-t, 0\} + \beta_i \max\{t-r, 0\}$$

as desired. △

$(2 \Rightarrow 1)$. R1-R4 are trivial. □

Proof of Proposition 4 $(1 \Rightarrow 2)$. By P1 $d_i(t,r) = d_i(0, r-t)$. By P2, whenever $r-t \leq 0$ we have $d_i(0, r-t) = 0$. When $r-t > 0$, by P3 it is true that $d_i(0, r-t) = (r-t)d_i(0,1)$. By P4 then $d_i(0, r-t) = r-t$. Thus, we can write $d_i(t,r) = \max\{r-t, 0\}$. △

$(2 \Rightarrow 1)$. P1-P4 hold trivially. □

Proof of Proposition 5 $(1 \Rightarrow 2)$. 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 P5 $D_i(x \,|\, \{x\}) = 0$ and by P6 $D_i(x \,|\, \{x, x_1\}) = d_i(u_i(x), u_i(x_1))$. Add elements one by one and use P6 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. △

$(2 \Rightarrow 1)$. P5 and P6 hold trivially. □

Proof of Proposition 6 $(1 \Rightarrow 2)$. For all $t_1, ..., t_N \in \mathbb{R}_+$ P2 implies $G(t_1, ..., t_N) = G(0, ..., 0) + \sum_{i \in N} \omega_i t_i$. By P1, $G(t_1, ..., t_N) = \sum_{i \in N} \omega_i t_i$. Thus, since D_i satisfy P5-P6

and d_i satisfy P1-P4, we have

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

as desired. △

$(2 \Rightarrow 1)$. P7-P8 are trivial. We get P1-P6 from the proofs of Propositions 4 and 5. □

Proof of Proposition 8 Consider a finite context $C \subset \mathcal{C}$ and two consequences $c_1, c_2 \in C$ with $u_i(c_1) \geq u_i(c_2)$ for all $i \in N$ with at least one strict inequality. For any i with $u_i(c_1) = u_i(c_2)$ we have $D_i(c_1|C) = D_i(c_2|C)$, and for any i with $u_i(c_1) > u_i(c_2)$ it is true that

$$D_i(c_1|C) = \sum_{c \in C} \max\{u_i(c) - u_i(c_1), 0\} < \sum_{c \in C} \max\{u_i(c) - u_i(c_2), 0\} = D_i(c_2|C).$$

Thus, $D(c_1|C) < D(c_2|C)$. The inequality is strict since $d_i(c_2, c_1) > 0$ for i with $u_i(c_1) > u_i(c_2)$. □

Proof of Proposition 10 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 11 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 8.

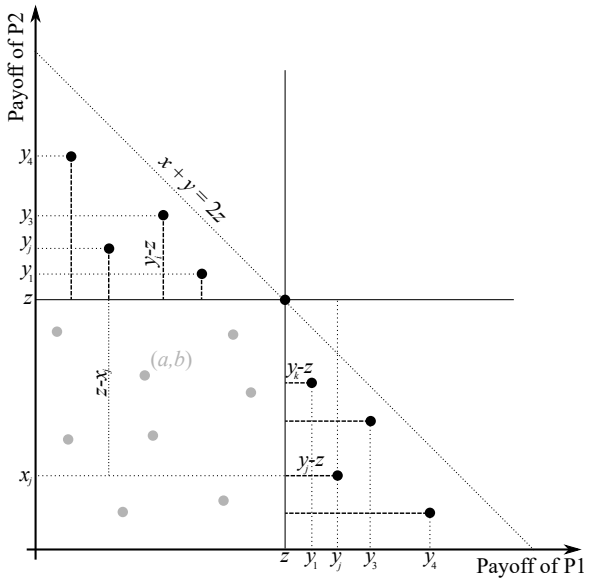

Figure 8: Illustration of the dissatisfaction calculations.

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 8 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 8). This makes (z, z) the P-norm. □

Proof of Proposition 12 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 P-norm by Proposition 8 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 11 are satisfied, which implies that (z, z) is the P-norm. □

Proof of Proposition 13 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. □

B Midpoint 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 P-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 P-model, it can be difficult to deal with such games due to the following. First, the P-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 utility vector $u[C]$ in $\mathbb{R}^N$ and its corresponding convex hull Ω, which is a convex *N-polytope*.[1] 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 6 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

[1] 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.

$(N-1)$-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.

B.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}.

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.

B.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 6, the P-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.[2]

B.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 6) 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 P-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 utility in Ω is z_i. Then, Figure 9 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 9, or it can be a continuum of allocations (the right panel). There are no intermediate cases since Ω is convex.

If z_i is achieved in only one allocation, then it is clear that there is high enough level of i's utility, say x_i, at which all allocations that give i utility x_i or more form a pyramid as on the left panel of Figure 9 (the pyramid is formed by the edges going

[2]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.

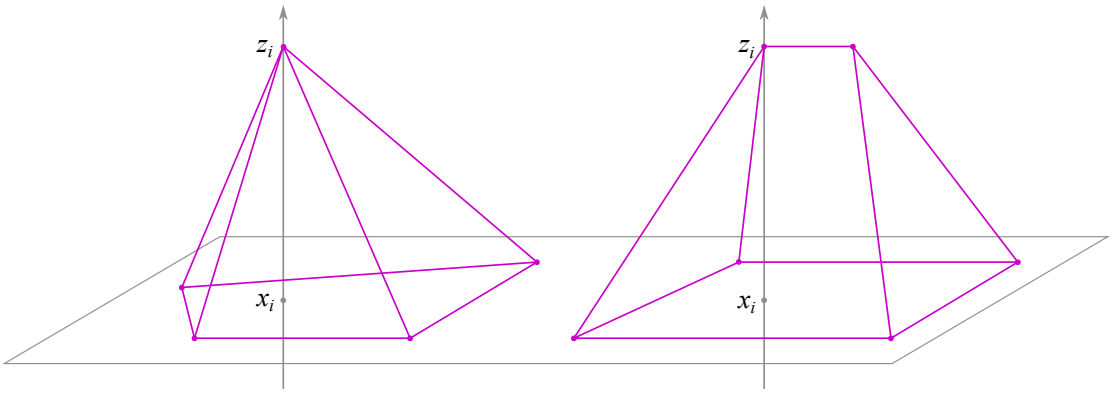

Figure 9: *Left.* The case when player i's highest utility z_i is achieved in only one allocation. *Right.* The case when player i's highest utility z_i is achieved in a continuum of allocations.

from the i's iso-utility plane containing x_i to the allocation that gives i utility 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 utility of player i higher or equal to x_i is formed by edges that go from the i's iso-utility plane containing x_i to some allocation that gives i utility 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 utility 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 utility in some allocation Z_i, then no other player can enjoy the highest utility at the same time. This expresses an idea that the utility in the game defined by Ω is somehow scarce for otherwise it would be possible to have more than one player enjoying the highest utility 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 utility 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.

B.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 utility z_i. We need this to prove our main result in the next section.

As we know from Section B.2, the personal dissatisfaction at allocation x that gives

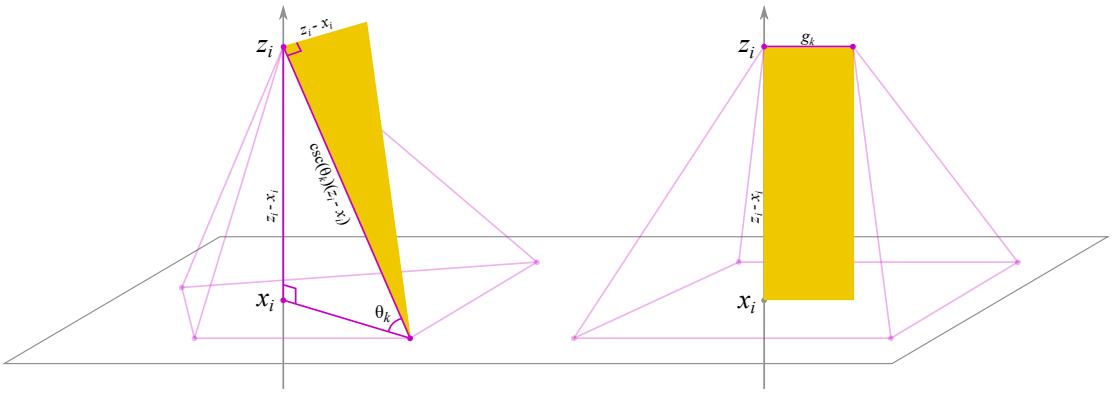

Figure 10: *Left.* A computation of the integral along the edge of the pyramid. *Right.* A computation of the integral along the horizontal edge with maximal utility z_i.

i utility x_i is the weighted sum of integrals of linear functions along the edges $\Omega^1_k(x_i)$ of the polytope $\Omega(x_i)$. Let us apply this result to the pyramid on the left panel of Figure 10 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-utility 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

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

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-utility 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 utility z_i:

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

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 $D_i(x \,|\, \Omega)$. Let us replace x_i with the variable $y_i = z_i - x_i$. Then, it is clear that $D_i(y \,|\, \Omega)$ 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 utility 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 14 *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 utility 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 10, 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-utility 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.

B.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 utility. In such situations allocating any

player the maximal utility 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 utility 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 utility is rather straightforward given the findings above. Indeed, take the allocation Z_i where i gets the maximal utility. 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 14 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 Midpoint Theorem to emphasize that in the normatively best outcome no player gets maximal utility.

Proposition 15 (Midpoint 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 10), 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 utility. 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 Midpoint Theorem presented above can be seen as a generalization of Proposition 10, 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.

C Experiment Details

References

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

Economic decision-making AXI2

Start of Block: Block 5

Prolific ID This is an experiment on economic decision-making. Please read the instructions carefully. You will answer four questions. After we collect data from 100 people, we will choose one of the questions randomly to be the "question that counts" for payment. We will pay you after the fact using your Prolific ID.

Please enter your Prolific ID. We need it to pay you. If the ID is incorrect you cannot be paid.

_____

End of Block: Block 5

Start of Block: DG

Allocation Decision You are deciding how to allocate money between two other people, call them Person A and Person B. If this question is chosen, we will randomly select two other survey respondents, and your answer will determine both people's payments.

Which of the following allocations of money do you prefer?

- ○ £0 for Person A and £3 for Person B (1)
- ○ £1 for Person A and £2 for Person B (2)
- ○ £2 for Person A and £1 for Person B (3)
- ○ £3 for Person A and £0 for Person B (4)

End of Block: DG

Start of Block: EG

Allocation Decision You are deciding how to allocate money between two other people, call them Person A and Person B. If this question is chosen, we will randomly select two other survey respondents, and your answer will determine both people's payments.

Which of the following allocations of money do you prefer?

- ○ £0 for Person A and £0 for Person B (1)
- ○ £1 for Person A and £1 for Person B (2)
- ○ £2 for Person A and £2 for Person B (3)
- ○ £3 for Person A and £3 for Person B (4)

End of Block: EG

Start of Block: Norms_DG

Appropriateness If this question is chosen, we will pick one of the rows below and compare your response to the most common response given by 100 people who answered this survey. If your response matches the most common response, you will receive £1. Otherwise you will receive £0.

Imagine a person is asked about these possible ways of allocating money between two strangers, call them Person A and Person B. We want to know how socially appropriate or how socially inappropriate it is to choose each allocation. For each row, please use the radio button to indicate how appropriate or inappropriate that action is. Remember, you get paid if your answer matches the most common answer given by other people.

	Very Socially Inappropriate (1)	Somewhat Socially Inappropriate (2)	Somewhat Socially Appropriate (3)	Very Socially Appropriate (4)
£0 for Person A and £3 for Person B (1)	○	○	○	○
£1 for Person A and £2 for Person B (2)	○	○	○	○
£2 for Person A and £1 for Person B (3)	○	○	○	○
£3 for Person A and £0 for Person B (4)	○	○	○	○

End of Block: Norms_DG

Start of Block: Norms_EG

Appropriateness If this question is chosen, we will pick one of the rows below and compare your response to the most common response given by 100 people who answered this survey. If your response matches the most common response, you will receive £1. Otherwise you will receive £0.

Imagine a person is asked about these possible ways of allocating money between two strangers, call them Person A and Person B. We want to know how socially appropriate or how socially inappropriate it is to choose each allocation. For each row, please use the radio button to indicate how appropriate or inappropriate that action is. Remember, you get paid if your answer matches the most common answer given by other people.

	Very Socially Inappropriate (1)	Somewhat Socially Inappropriate (2)	Somewhat Socially Appropriate (3)	Very Socially Appropriate (4)
£0 for Person A and £0 for Person B (1)	○	○	○	○
£1 for Person A and £1 for Person B (2)	○	○	○	○
£2 for Person A and £2 for Person B (3)	○	○	○	○
£3 for Person A and £3 for Person B (4)	○	○	○	○

End of Block: Norms_EG