Injunctive Norms and Moral Rules[*]

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

January 2020

Abstract

We show how the psychological notion of dissatisfaction can be used to build an axiomatic model of choice-set-dependent injunctive norms. In an accompanying paper (Kimbrough and Vostroknutov, 2020) we demonstrate that a functional representation obtained from these axioms has substantial explanatory power in many types of experimental games. Therefore, the axioms provide a realistic description of a class of models of injunctive norms that can be used to predict social behavior. Nevertheless, it follows from the axioms that the relative normative valences of any fixed outcomes can be highly dependent on the other options available in the choice set. This makes such norms hard to compute, and may drive people to using moral rules instead of injunctive norms. A moral rule is a simple but relatively good approximation of normative valences in a specific class of choice settings. With another set of axioms we show how to define choice-set-dependent moral rules like Pareto optimality, payoff efficiency, maximin, or inequality aversion. We compute the predictive power of these approximations in various classes of choice settings. This methodology allows one to calculate which moral rule is likely to be used in some specific strategic situation, which is easily testable and can help to significantly improve our understanding of incentivized social behavior.

JEL classifications: C91, C92
Keywords: norms, axiomatic preferences, rules.

[*] We would like to thank Elias Tsakas for help with formulating the axioms. 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

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, individuals are variously motivated by Pareto improvements, efficiency, equality, maximin, reciprocity, guilt aversion, lying aversion, anger, and so on. Moreover, 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). What determines the mapping between context and the choice criterion? A promising approach to a unifying explanation argues that apparent context-dependence in motivation reflects adherence to context-dependent injunctive social norms (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 why and how social norms vary across contexts.

To address this gap in the literature, we build an axiomatic model of choice-set-dependent injunctive norms. We construct norms from the ground up, starting with axioms on the psychology of normative judgment and then providing additional axioms on the process of aggregation that transforms individual normative judgments into broadly shared injunctive norms. The model begins with the assumption that normative judgments reflect a basic aspect of human psychology: that at any outcome people are dissatisfied with *what they have* because of *what they might have had instead.* This is the sense in which normative judgments are choice-set-dependent; each person's normative evaluation of each outcome depends directly on how it compares to all other possible outcomes. An outcome is more dissatisfying the more numerous (and better) are the other available options that would be preferable to it. We combine this assumption with another basic human psychological ability: the capacity for empathy, that is, 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 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. Thus injunctive norms inherit the property of choice-set-dependence.

In a companion paper (Kimbrough and Vostroknutov, 2020), we show that a functional representation obtained from our 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 injunctive norms. Introducing choice-set-dependence thus allows us to account for a number of otherwise puzzling observations, but it comes with a cost: as the complexity of the choice setting grows, identifying the norm involves increasingly strong assumptions about one's knowledge of others' preferences and increasingly complex computations. While choice-set-dependent norms of the kind modeled here seem to track how people's normative intuitions and their behavior vary with context in laboratory settings, for practical purposes (e.g. in organizations, in the law) such norms are difficult to identify, specify, and articulate.

By way of contrast, we introduce an alternative set of axioms that are not so radically choice-set-dependent and show how norms that are consistent with many of the widely-used models of social behavior (e.g., social preference models) can be derived from these axioms by introducing an ideal reference outcome and computing dissatisfaction relative to that ideal. Norms derived from the second set of axioms can usually be summarized via general principles like those mentioned above (e.g. maximin, payoff efficiency), but these theories have the drawback that *any* such theory cannot accommodate numerous experimental observations where behavior is consistent with different general principles in different contexts. That is, such theories are not truly context-dependent since they involve imposing a pre-conceived rule, constant across at least some contexts, and we show that the normative evaluation of outcomes implied by these axioms are not equivalent to those generated by our choice-set-dependent axioms. The implication is that it is impossible to summarize norms of the kind implied by our first set of axioms with simple, general principles.

Despite this impossibility result, our analysis suggests a deep connection between the norms implied by our theory and the kinds of general principles that have been offered as guides to behavior (and as theories of social preferences). In our view, the complexity of identifying and implementing norms that satisfy the choice-set-dependent axioms induces a search for lower-complexity substitutes in the form of general principles that approximate these norms. We view most models of social preferences and social norms as capturing species of such approximations; by introducing an ideal reference outcome and computing dissatisfaction relative to that ideal, it is possible to approximate the kinds of judgments that emerge from the more complicated choice-set-dependent axioms in a wide variety of settings. We refer to such approximations as *moral rules*.

We show how moral rules can arise naturally by identifying salient properties of normative judgments that are frequently implied by our first set of axioms and then generalizing them across contexts. Thus, moral rules can be thought of as heuristics, summarizing common normative judgments that arise from the choice-set-dependent axioms, rendering them easily communicated, and facilitating coordination. For example, our theory implies that in particular contexts, each of the following three moral rules would accurately characterize the normatively appropriate action:

- *The efficiency rule* - always prefer the outcome with the highest sum of utilities
- *The maximin rule* - always prefer the outcome that maximizes the minimum utility
- *The Pareto rule* - always prefer Pareto improvements

For instance, in choice sets with only two outcomes, our choice-set-dependent theory is perfectly consistent with the efficiency rule (under linear utility) or the maximin rule (under sufficiently concave utility). More generally, imposing various restrictions on the set of possible outcomes (e.g., random payoff vectors, constant-sum payoff vectors, social dilemmas, coordination games, etc.) we can ask which moral rule is most consistent with the norms implied by our choice-set-dependent axioms and thereby predict the kinds of rules people would be expected to articulate if explaining appropriate behavior to one another in settings of that kind. That is, our framework provides a meta-theory of moral rules, showing their origins (as approximations of the norms that emerge from our choice-set-dependent axioms), and how to find out which rules are likely to emerge in which contexts.

2 Dissatisfaction Functions for Injunctive Norms

The theory begins with the assumption that an agent's own normative evaluation of any particular outcome depends on how it compares to all other feasible outcomes. Put simply, agents prefer outcomes that yield higher utilities and are dissatisfied with outcomes that yield lower utilities *because of* the counterfactual availability of higher-utility outcomes. First we define a function that captures this notion of dissatisfaction.[1]

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 $\mathcal{C}$, with a typical one called $C \subset \mathcal{C}$. Analogously to Kimbrough and Vostroknutov (2020), we think of C and the collection of associated payoff vectors $u[C]$ as a set of all feasible allocations in a choice problem. In addition, suppose that there are functions $d_i : \mathbb{R}^2 \to \mathbb{R}_+$ that for any consequences $x, y \in \mathcal{C}$ and player i define i's *dissatisfaction with x because of y* as $d_i(u_i(x), u_i(y))$. Assume that for all $i \in N$, $d_i(u_i(x); u_i(y))$ is weakly increasing in $u_i(y)$, which represents the idea that counterfactual outcomes with higher utility increase dissatisfaction.

Given these ingredients, in the first step we define dissatisfaction functions $D_i(c \,|\, C)$ for each player i that for each subset C and each consequence $c \in C$ gives a non-negative real number representing the *total dissatisfaction* associated with c in C of player i. This provides us with a theory of individuals' normative evaluations of each outcome, as a function of all other possible outcomes - a choice-set-dependent model of normative judgment.

In the second step we define an *aggregate dissatisfaction function* $D(c \,|\, C)$ that creates a composite of the dissatisfaction of all interested players. This function then directly translates into an injunctive normative valence associated with an element c given the set of possibilities C. We propose a set of axioms that describe the properties of d_i, D_i, and D. We start with the axioms that define the connection between dissatisfaction d_i and utility of player i.

A1 (Relativity). $\forall i \in N \; \forall t, r, a \in \mathbb{R} \quad d_i(t + a, r + a) = d_i(t, r)$.

A1 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 this does not affect her dissatisfaction. A1 ensures that d_i can be expressed as a function of the difference of utilities.

A2 (No Rejoice). $\forall i \in N \; \forall t, r \in \mathbb{R}$ with $t \geq r$ we have $d_i(t, r) = 0$.

[1] Conceptually, these normative evaluations are made prospectively, before an outcome has been realized. We suggest in Kimbrough and Vostroknutov (2020) that an appropriate moniker for the sentiment being captured might be "pre-gret".

A2 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. It is important to note that we do not assume here that players do not feel any such sentiment in these circumstances, only that this sentiment does not influence the normative valence of superior outcomes.

A3 (Homogeneity). $\forall i \in N\ \forall t, r, \alpha \in \mathbb{R}$ with $\alpha > 0 \quad d_i(\alpha t, \alpha r) = \alpha d_i(t, r)$.

A3 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 that there is some non-linear, say concave, relationship between dissatisfaction and utility. However, we already allow utility to be a non-linear function of payoffs. A3 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.

A4 (Non-triviality and Anonymity). $\forall i \in N \quad d_i(0, 1) = 1$.

A4 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 is 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 normative evaluations are built on empathy.

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

Proposition 1. The following two statements are equivalent:

1. d_i satisfies A1-A4*;*
2. $d_i(t, r) = \max\{r - t, 0\}$.

Proof. See Appendix A.

Next, we provide axioms that define the total dissatisfaction D_i of player i associated with a single consequence c.

A5 (Singleton). $\forall i \in N\ \forall c \in \mathcal{C} \quad D_i(c \mid \{c\}) = 0$.

Axiom A5 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. A5 may sound trivial, nevertheless it rules out any situations in which players are dissatisfied due to specific properties of an allocation c given the choice set $\{c\}$. This makes our model incompatible with social preference utility specifications where the utility of a player may depend directly on the payoffs received by other players at the same consequence. A form of social preferences, in the common meaning of the term, could be introduced if in A5 we assumed that dissatisfaction is not zero, but rather depends on c. 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, 2020)). Indeed, our goal 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.

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

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

Axiom A6 defines the total 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, which is guaranteed by the assumptions on d_i made 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 it 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 A6 as capturing another basic sentiment, namely that player i feels dissatisfaction whenever a new possibility represented by y that could give i a higher payoff appears.

The following result connects the axioms above and the representation that we test in Kimbrough and Vostroknutov (2020).

Proposition 2. The following two statements are equivalent:

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

Proof. See Appendix A.

Finally, we aggregate the dissatisfactions across players and define D. We start by assuming that $D(c \,|\, C)$ is a function of $D_i(c \,|\, C)$ for all $i \in N$. Specifically, that $D(c \,|\, C) = G(D_1(c \,|\, C), ..., D_N(c \,|\, C))$, where $G : \mathbb{R}^N \to \mathbb{R}_+$ is increasing in all arguments. The following axioms determine how aggregation is done.

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

A7 simply states that if each player feels the lowest dissatisfaction of zero, then the aggregate dissatisfaction is also the lowest and equals to zero.

The last axiom defines how changing dissatisfaction of one player changes the aggregate dissatisfaction. In order to incorporate social context as defined in Kimbrough and Vostroknutov (2020), we assume that players have social weights $(\omega_i)_{i \in N}$ where $\omega_i \in (0, 1]$. These weights can represent social status, in/outgroup relationships, kinship, or their combination and they determine how much each player's dissatisfaction counts in the computation of aggregate dissatisfaction.

A8 (Aggregate Change). $\forall i \in N\ \forall t_1, ..., t_N \in \mathbb{R}_+\ \forall a_i \geq -t_i \quad G(t_i + a_i; t_{-i}) = G(t_i; t_{-i}) + \omega_i a_i$.

The notation $G(t_i; t_{-i})$ singles out the ith argument of G. A8 says that if player i's total dissatisfaction changes by a_i then the aggregate dissatisfaction changes by the same amount weighted by ω_i. This incorporates an idea that social 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. d_i satisfies A1-A4, D_i satisfies A5-A6, D satisfies A7-A8.
2. D can be expressed as

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

Proof. See Appendix A.

In Kimbrough and Vostroknutov (2020), we show how to use the representation in Proposition 3 in the norm-dependent utility function that combines material payoffs and the desire to follow injunctive norms.

3 Dissatisfaction Functions for Moral Rules

In this section we describe an alternative 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; this might include concepts like payoff efficiency, Pareto optimality, equality, or some combination of these, or any other *abstract* 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. These axioms serve several purposes. First, they show how to represent *any* choice-set-dependent moral rule, which defines the normatively best element of all choice sets C, with a dissatisfaction function. Second, contrasting the dissatisfaction functions representing these abstract moral rules with the radically choice-set-dependent dissatisfaction functions described above, we highlight the extent to which the former can serve as approximations of the normative evaluations implied by the latter.

The key insight is that the normative evaluations captured in the dissatisfaction functions that result from our first set of axioms cannot be replicated by a moral rule. This is important because we often communicate our moral arguments in terms of such rules, and we often want to codify normative judgments into laws or other forms that must be explicitly stated (e.g., religious texts). The issue is that abstract ideals are necessarily simplifications and thus they are unable to capture some of the richness and complexity that arises from our model of injunctive norms. That is, there will always be implications from the radically choice-set-dependent axioms that will not follow from an abstract moral rule. This result demonstrates the possibility that there may not exist an ideal-coded set of explicit rules that exactly capture intuitive normative perceptions, and may imply that situations in which these normative perceptions are in conflict with codified normative ideals will always arise.

As before, we start with a large set $\mathcal{C}$ of all consequences, its finite subsets C, and a utility function $u : \mathcal{C} \to \mathbb{R}^N$. The difference from the previous analysis is that now 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 $u^*(r \mid S)$ that for each vector $r \in \mathbb{R}^N$ and all subsets $S \subset \mathbb{R}^N$ that are equal to $u[C^*]$ for some C defines an element in S that is a "reference point" for r in S.[2] 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: $u^*(r \mid S) = r$ whenever $r \in S$, which makes sure that 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 dissatisfactions for each player $i \in N$. Namely, $f_i(r_i, u_i^*(r \mid S))$ stands for the dissatisfaction that player i feels when her utility is r_i and the utility that she would receive in the ideal situation is $u_i^*(r \mid S)$.

In what follows, we denote the second set of axioms as B1 …B7, for ease of exposition we distinguish between the A-model, which captures our radically choice-set-dependent theory of normative evaluations presented above, and the B-model, which includes an ideal. As above, to describe the B-model we start with the axioms describing the properties of f_i.

B1 (Relativity). $\forall i \in N\ \forall t, r, a \in \mathbb{R} \quad f_i(t + a, r + a) = f_i(t, r)$.

B1 is exactly the same as A1, and it states that adding a constant to the utilities being compared does not change the dissatisfaction. So, only the difference in utilities matters.

B2 (No Rejoice). $\forall i \in N\ \exists \beta_i \in (0, 1]$ such that $\forall t \geq 0$ we have $f_i(0, -t) = \beta_i f_i(0, t)$.

B2 is different from A2. It states that there might be an asymmetry in how dissatisfaction is perceived when the ideal utility is above or below the received one. 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 overall dissatisfaction is zero (all players

[2] Specifically, there is no need to define $u^*(r \mid S)$ 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. In there we need to define u^* for only one set S.

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^*.

B3 (Homogeneity). $\forall i \in N\ \forall t, r, \alpha \in \mathbb{R}$ with $\alpha > 0 \quad f_i(\alpha t, \alpha r) = \alpha f_i(t, r)$.

This axiom is the same as A3. Finally, as above we assume non-triviality and equivalence of dissatisfactions across players.

B4 (Non-triviality and Anonymity). $\forall i \in N \quad f_i(0, 1) = 1$.

The following proposition provides the representation.

Proposition 4. The following two statements are equivalent:

1. f_i satisfies B1-B4*;*
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 total dissatisfaction $F_i(c \,|\, C)$ of player i. This is very straightforward. However, it is worth noticing before we state the axiom that unlike in the A-model, in the B-model we do not require that F_i is zero for singleton choice sets $C = \{c\}$. This is because it is easy to imagine a situation where there is only one allocation which nevertheless deviates from the ideal. For example, when the ideal for each player is the average payoff in c (see Section 3.1). This highlights the sense in which ideals involve abstraction; in the A-model this is not possible since the normatively best consequence is defined endogenously from the available consequences in C.

B5 (Dissatisfaction). $\forall i \in N\ \forall C \subset \mathcal{C}, x \in C \quad F_i(x \,|\, C) = f_i(u_i(x), u_i^*(u(x) \,|\, u[C^*]))$.

In words, player i's total 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, B5 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, the i's total dissatisfaction of all elements $x \in C_1 \cap C_2$ is the same. This is the main feature that distinguishes the B-model from the A-model, where such connection is only attainable in specific rare cases. 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 all elements of these sets are used to compute them. So, instead of providing a representation result, we formulate this implication as a proposition.

Proposition 5. *B5 implies that for any C_1 and C_2 with $u[C_1^*] = u[C_2^*]$ it is true that $F_i(x \,|\, C_1) = F_i(x \,|\, C_2)$ for all $x \in C_1 \cap C_2$.*

Proof of Proposition 5. By the property of u^*. ■

Finally, we define aggregated dissatisfaction $F(c \,|\, C)$. This is done in the same way as in the A-model.

We assume that F is a function of F_i's: $F(c\,|\,C) = E(F_1(c\,|\,C), ..., F_N(c\,|\,C)$. The following two axioms describe the properties of E.

B6 (Aggregate Satisfaction). $E(0, ..., 0) = 0$.

This is the same as A7. The last axiom is the same as A8 assuming the existence of social weights $(\omega_i)_{i \in N}$.

B7 (Aggregate Change). $\forall i \in N\ \forall t_1, ..., t_N \in \mathbb{R}_+\ \forall a_i \geq -t_i \quad E(t_i + a_i; t_{-i}) = E(t_i; t_{-i}) + \omega_i a_i$.

The following proposition puts all B-axioms together.

Proposition 6. The following two statements are equivalent:

1. f_i satisfies B1-B4, F_i satisfies B5, F satisfies B6-B7.
2. F can be expressed as

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

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

Proof of Proposition 6. Similar to the proof of Proposition 3. ■

The following examples show how to express some commonly studied moral rules using a dissatisfaction function F that satisfies the axioms above.

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 not empty for all C. $u^*(r\,|\,S)$ can be defined as an element of S that is the closest to r in Euclidean metric.

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 $y \notin C^*$. $u^*(r\,|\,S)$ 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 $y \notin C^*$. $u^*(r\,|\,S)$ 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^*.[3]

[3] 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.

3.1 Dissatisfaction Functions for Choice-Set-Independent Moral Rules

It is worth paying attention to the simplest class of ideal dissatisfaction functions for which the ideal set C^* is independent of C. We assume that for all $C \in \mathcal{C}$ there is only one C^*, and that the B-axioms hold as before. In this case for any non-ideal $c \in \mathcal{C}$ the dissatisfaction from c is the same for all $C \in \mathcal{C}$ that include c. In other words, $F_i(c \,|\, C) = F_i(c) = f_i(u_i(c), u^*(u(c) \,|\, u[C^*]))$. This model closely approximates outcome-based social preference models, with the only difference being that in our approach all players have the same overall dissatisfaction $F(c)$ attached to an outcome, whereas in standard 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). One argument in favor of our way of modeling social preferences is that the construction of the norm from aggregate (and not only individual) dissatisfaction, since $F(c)$ is interpreted as the aggregate dissatisfaction of *all* players, means that we prioritize the *social* component of social preferences. In our framework, each player cares about all others, since their dissatisfaction enters her utility function through the norm, and not only about how much their own payoff diverges from that of others.

The following example illustrates what a choice-set-independent dissatisfaction function may look like.

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

4 Non-Equivalence of Injunctive Norms and Moral Rules

In this section we show that abstract moral rules of the kind captured in the B-model cannot represent the complexity of the norms generated by the A-model.[4] To do this we need a notion of equivalence between them. Given that the d and f functions in A- and B-models are not fixed, we can 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 dissatisfactions in the A-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 B-model defined by the collection of all C and C^* together with a distance function f let $\dot{\succcurlyeq}_C$ represent the dissatisfactions:

$$\forall C \in \mathcal{C} \quad x \;\dot{\succcurlyeq}_C\; y \Leftrightarrow F(x \,|\, C) \geq F(y \,|\, C).$$

Let us say that a B-model is *equivalent* to the A-model if $\succcurlyeq_C$ and $\dot{\succcurlyeq}_C$ are the same for all $C \in \mathcal{C}$. Our task now is to show that there is no B-model that is equivalent to the A-model satisfying all A-axioms.

[4] Of course, ideal models can generate many things that A-model cannot, freely defined essentially by all C and C^*.

To do that, let us try to construct a B-model that is as close as possible to the A-model. B-models are defined through ideal elements C^* and a distance function. So, for each C define C^* as the set of minimal elements of $\succcurlyeq_C$ for each C and choose any dissatisfaction function f.

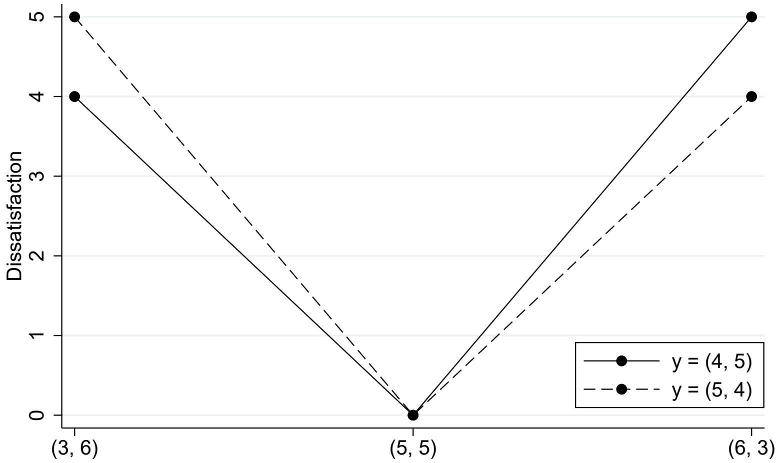

Figure 1: A-model dissatisfactions of the same allocations in $C_y = \{(3,6);(5,5);(6,3);y\}$, where $y = (4,5)$ or $y = (5,4)$.

The following example shows how no such model can be equivalent to the A-model. Consider sets of four allocations for two players $C_y = \{(3,6);(5,5);(6,3);y\}$ where y can take on different values, say $y = (4,5)$ or $y = (5,4)$. In the A-model both $C_{(4,5)}$ and $C_{(5,4)}$ have minimal dissatisfaction at $(5,5)$. So, when we construct a B-model as described above for both of them we set $C^* = \{(5,5)\}$. By Propositions 5 and 6, in any B-model $F((3,6)\,|\,C_{(4,5)}) = F((3,6)\,|\,C_{(5,4)})$ and the same holds for allocation $(6,3)$. Thus, according to any B-model either $(3,6)$ or $(6,3)$ has *larger* dissatisfaction in *both* $C_{(4,5)}$ *and* $C_{(5,4)}$. However, Figure 1 shows that this is inconsistent with the A-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 B-model is equivalent to the A-model.[5] We state this result as a proposition.[6]

Proposition 7. *There is no B-model that is equivalent to the A-model in the sense of equivalence of the preference orderings of dissatisfactions that the models imply.*

Proof of Proposition 7. By example on Figure 1. ■

This proposition demonstrates that if it is the aggregate dissatisfaction that people care about—as the

[5] It is important to note at this point that the dissatisfactions of the payoff vectors which are not in C^* or not the A-norm may matter for many reasons. For example, since we always assume that players solve the trade-off between personal consumption and norm-following (Kimbrough and Vostroknutov, 2020), the dissatisfactions of all payoff vectors become relevant for norm-dependent utility maximization. Moreover, the dissatisfactions of non-normatively best consequences are used to compute the severity of norm violation used in punishment. Thus, having an ideal with all sets C^* coinciding with the A-norm is not enough.

[6] It may seem that we could construct an equivalent B-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.

A-model postulates—then there is no B-model that reflects the same criterion. We have shown that this is the case with the class of B-models that are the closest to the A-model, namely those that have the same minimal dissatisfaction as the A-model in all sets C. For more general B-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 A-model in general the dissatisfactions of all elements will change after such addition, and the minimal dissatisfaction will not be assigned in the same way by the A- and B-models. Moreover, the dissatisfaction-minimizing allocations in the A-model are never Pareto-dominated (see Proposition 1 in Kimbrough and Vostroknutov, 2020). Thus, if in a B-model some C^* contains Pareto-dominated consequences, as for example in the Inequality Aversion examples described above, then this will always contradict the predictions of the A-model except for some special cases.

5 Why Moral Rules?

As shown above, the radical choice-set-dependence of the A-model of injunctive norms, in which the evaluation of each consequence depends on all other feasible consequences, makes it impossible to characterize its normative implications succinctly with an abstract ideal. If we take seriously the idea that the A-model provides an account of peoples' normative intuitions, this poses a problem - it will not, in general, be possible to specify abstract ideals whose implications will satisfy our normative intuitions in all cases. Dissatisfaction will not in general be minimized by following moral rules. What good, then, are abstract moral rules? We argue that one might nevertheless prefer an inexact moral rule to a precise injunctive norm because of the latter's relative complexity.

Consider a simple case with the social weights of all players equal to 1. Suppose that there are N players and consider some set of consequences C that has k elements. Then, to compute normative valences of all elements of C in the A-model one needs to compute Nk^2 payoff differences (and decide whether they are greater or equal to zero). So, we can say that the complexity of this problem is $O(Nk^2)$. For the similar B-model with all $\beta_i = 1$ one needs to compute only Nk payoff differences and decide whether they are greater or less than zero. So the complexity of this problem is $O(Nk)$. So, for any fixed number of players, the normative valences in the A-model can be computed in quadratic time, whereas the valences in any B-model can be computed in linear time. As the number of consequences grows, so does the appeal of an abstract rule for assigning normative valences. Of course, this ignores the computations involved in finding C^* for each C, but here too moral rules would seem to have an advantage because they can be articulated in a way that generalizes across choice settings (e.g., "one ought to maximize efficiency"), facilitating coordination.

This is an argument for why we might wish to have some clearly articulated abstract moral rules which we use to guide behavior, but it leaves open the question of what those moral rules might look like. We suggest that the most appealing moral rules for a given class of choice settings would be those which most often approximate the normative intuitions captured in injunctive norms computed in those settings. We thus ask, for various classes of choice settings, what are the commonly observed properties of the injunctive norm generated by the A-model in those settings that might be summarized as a moral rule. First, we show that regardless of the choice setting, all dissatisfaction-minimizing outcomes under the A-model will be Pareto optimal, such that this moral rule is always satisfied by an injunctive norm. Second, for the special case of choice settings with exactly two possible consequences, we show that the injunctive norm implies payoff efficiency maximization. More generally, using simulations in which we append random payoff vectors to various numbers of consequences under different restrictions (e.g., constant sum), we ask how frequently the injunctive norm, calculated using the A-model, corresponds to various other plausible (or commonly expressed) moral rules: "maximize efficiency", "minimize inequality", etc. Then we consider the moral rules that might be derived from injunctive norms for payoff vectors that correspond to broad classes of games that are of interest to economists, such as social dilemmas, coordination games and tournaments. We show how our A-model can be used to provide a meta-theory of moral rules, predicting which moral rules are likely to emerge in which choice settings. As it turns out, the injunctive norms implied by the A-axioms often cohere nicely with existing moral rules such efficiency maximization, inequality aversion, etc.

5.1 Moral Rules for All Choice Settings

As we have shown in Kimbrough and Vostroknutov (2020), the injunctive norms implied by the A-model have an important property: in *any* set of consequences, a Pareto dominated consequence is always normatively inferior to the one that Pareto dominates it (i.e., it always generates more aggregate dissatisfaction). This regularity is easy to spot from observation. Thus injunctive norms derived from the A-model always satisfy the Pareto optimality criterion. This is the only moral rule that is exactly represented by the A-model; however, it is worth noting that our model goes further in that it also provides 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 A-model is limited.

5.2 Moral Rules for Random Payoff Vectors

First, from Kimbrough and Vostroknutov (2020) we know that, as long as agents' utility functions do not exhibit too much curvature, payoff efficient allocations are dissatisfaction-minimizing in the A-model for all choice sets with exactly two consequences. For larger sets of consequences, this is not true in general. Nonetheless, it is true quite frequently, enough so that it is plausibly reasonable to summarize this tendency of the injunctive norm as a moral rule. To illustrate the point, we simulate random sets of payoff vectors of varying sizes and for different numbers of players, and compute the percentage of cases in which the dissatisfaction-minimizing consequence under the A-model (hereafter, A-norm) would also be chosen by the moral rule "choose the most efficient outcome."[7]

[7] For each given number of players and given number of consequences, we generate 500,000 sets of random payoff vectors with payoffs independently drawn from the uniform distribution on $[0, 1]$.

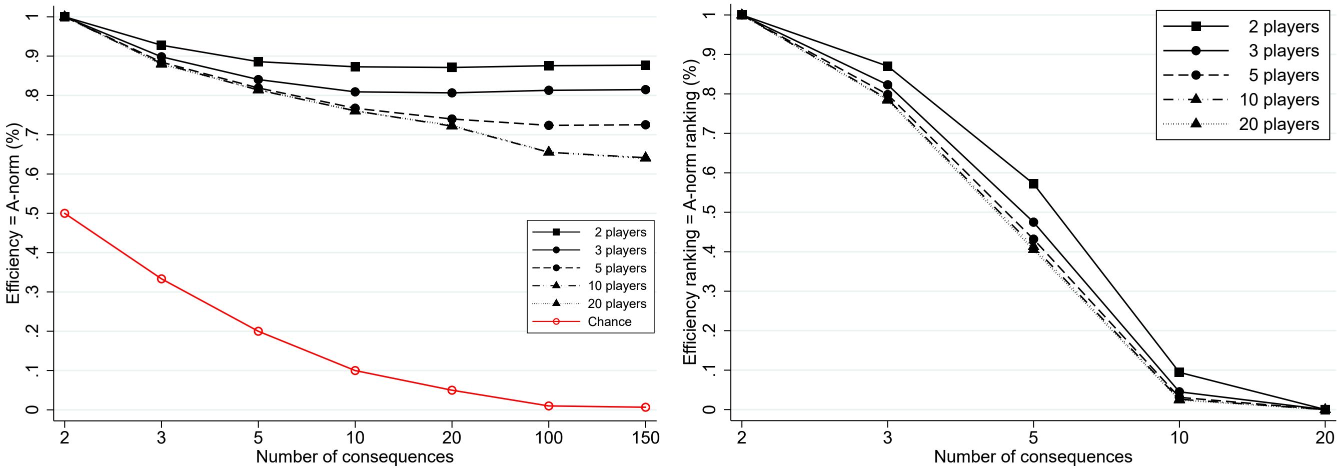

Figure 2: Left: the percentage of A-norms that are payoff efficient for random sets of consequences. 500,000 sets for each case. Right: the percentage of cases in which the A-dissatisfaction function ranks outcomes according to their payoff efficiency.

The left panel of Figure 2 shows the results. With 3 players and a large number of consequences, around 80% of A-norms are payoff efficient. This percentage tails off as the number of players grows. With 10 or 20 players and large sets of consequences the number of payoff efficient A-norms drops to 65%, but the correspondence remains striking, happening far more often than would be predicted by chance alone. A more stringent test asks whether the ranking of outcomes according to the A-model is identical to the ranking according to payoff efficiency, or what percent of the time the A-model is equivalent to the moral rule in the sense defined in Section 4. The right panel of Figure 2 shows that the number of equivalent rankings goes to zero rather rapidly with the number of consequences.

This simple analysis shows the value of our framework. For example, we can hypothesize that the payoff efficiency criterion may be a good approximation of the default moral rule for small sets of consequences and players. As these numbers grow, payoff efficiency becomes progressively worse at predicting the most normatively desirable outcome (A-norm). This suggests that some other moral rule might begin to look appealing in such cases. We summarize this as a conjecture.

Conjecture 1. *For small sets of randomly chosen payoff vectors with 3 to 5 consequences, people will identify maximization of payoff efficiency as a moral rule since it coincides with the A-norm in 80-90% of cases. As the set of consequences and the set of players grow, the payoff efficiency heuristic should be used less often.*

Next we consider maximin criterion as a moral rule for random payoff vectors (see the end of Section 3 for the definition). The left panel of Figure 3 shows the percentages of cases in which the A-norm coincides with the maximin ideal. Overall maximin rule corresponds to the A-norm less often than payoff efficiency. However, the difference becomes considerable only for large sets of players and large sets of consequences. Therefore, we cannot a priori rule out the possibility that maximin can be used as a moral rule for small numbers of players and small sets of consequences. Indeed, Engelmann and Strobel (2004) and Baader and Vostroknutov (2017) show that around 40% of subjects choose according to maximin in 3-player mini-dictator games with 3 consequences. Interestingly, other 40% follow payoff efficiency criterion, which also predicts the A-norm very well in such situations: for 3 players and 3 consequences the number of payoff efficient A-norms is 90% and the number of maximin A-norms is around 75% (see Figures 2 and 3).

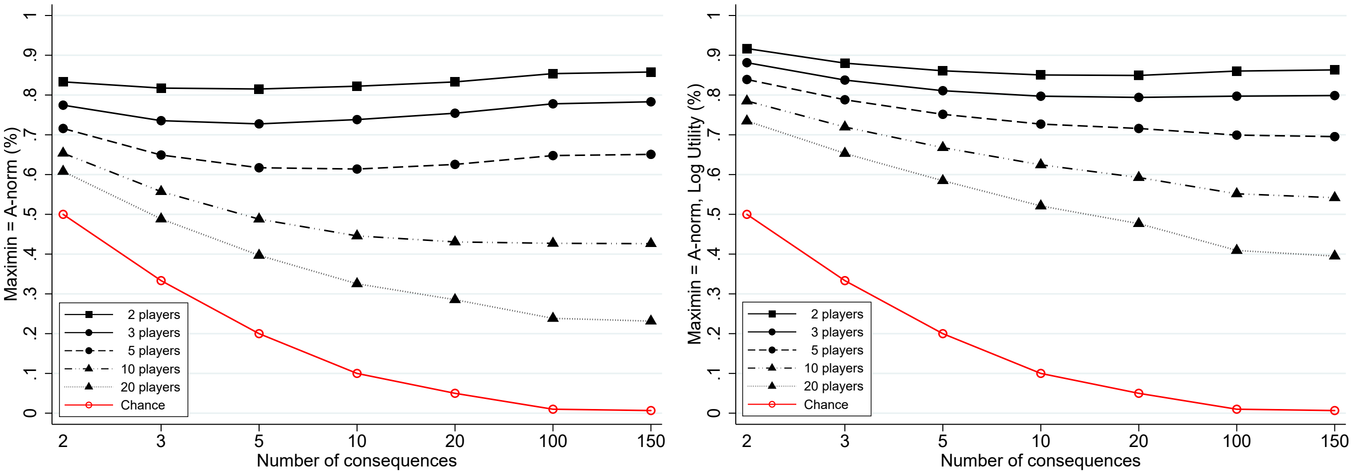

Figure 3: Left: the percentage of A-norms that are maximin for random sets of consequences. Right: the percentage of A-norms with log utility that are maximin for random sets of consequences. 500,000 sets for each case.

In Kimbrough and Vostroknutov (2020) we argue that maximin is more likely to be the A-norm when players use log utility to compute dissatisfactions. The reason is that log utility embellishes the dissatisfactions of "poor" players, in the sense that the same logged-payoff difference generates much more dissatisfaction of poor than rich players. Thus, we argue that people who are prone to compute dissatisfactions taking relative wealth of others into account are more likely to follow maximin as a moral rule (for some extreme examples of such behavior see MacFarquhar, 2016). The right panel of Figure 3 shows the percentages of cases in which the A-norm coincides with the maximin rule in randomly generated logged-payoff vectors. The performance of the maximin rule increases—as compared to the left panel of Figure 3—especially for the large sets of players and consequences. This provides support to our intuition in Kimbrough and Vostroknutov (2020). We summarize our findings as a conjecture.

Conjecture 2. *For small sets of players (2 to 5) and randomly chosen payoff vectors with any number of consequences, people will identify maximin as a moral rule since it coincides with the A-norm in 70-90% of cases under the assumption that people use log utility to compute dissatisfactions. This also holds to a lesser degree even without log utility. As the set of players grows, the maximin rule should be used less often. Given that both payoff efficiency and maximin fit the A-norm rather well when the set of players is small, we should expect to observe both rules used in these circumstances.*

By way of contrast, we can also ask whether other well-known moral rules correspond to the A-norm in these settings. For example, if we repeat the exact same exercise as above and ask how often the A-norm makes the same prediction as inequality aversion in randomly chosen payoff vectors, we get

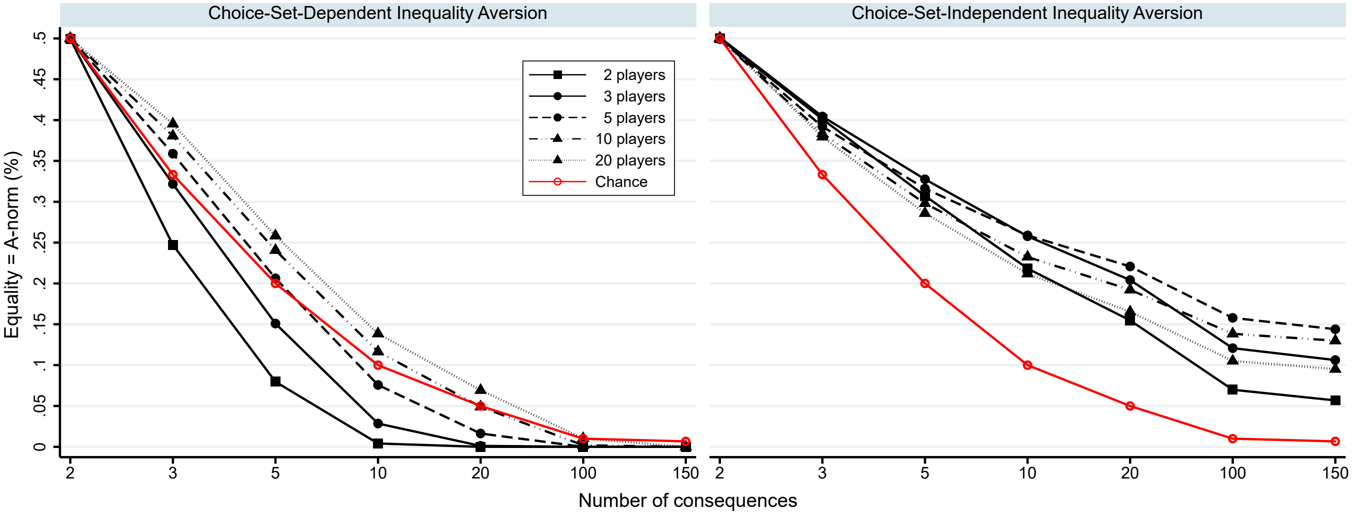

Figure 4: The percentage of A-norms that coincide with predictions of Choice-Set-Dependent and Choice-Set-Independent Inequality Aversion for random sets of payoff vectors. 500,000 sets for each case.

rather different results. Figure 4 shows that for general random sets of payoff vectors, the choice-set-dependent inequality aversion model corresponds to the A-norm at a rate no better than chance, while the choice-set-independent model does only slightly better than chance. This is not surprising given that the A-norm is always Pareto optimal, whereas inequality aversion models tend to favor payoff vectors close to the center of any set C. However, it is plausible that inequality aversion models may be better suited to situations like the Dictator game in which payoff efficiency is constant and the payoffs of each player are non-negative. We explore the properties of the A-norm in those settings next.

5.3 Moral Rules in Constant-Sum Settings

In the above example, we computed injunctive norms over randomly chosen payoff vectors. A random set of payoff vectors has with probability 1 a unique highest efficiency element, which is often favored by the A-model. So, the 80% result above should be taken with care.[8] Moreover, the sets of payoff vectors reside in a multidimensional space, and it is easy to find a measure zero subspace which corresponds to a particular widely-studied game and for which the payoff efficiency moral rule may not correspond to the A-norm. For example, the set of payoff vectors corresponding to a Dictator game has measure zero in the two-dimensional space of payoff vectors for 2 players. The results in Figure 2 cannot possibly be influenced by the properties of such measure zero subspaces since the probability that the sets of payoff vectors from these subsets are randomly drawn is zero. That is, if we randomly draw sets of payoff vectors from $\mathbb{R}^2$ we will never draw one where all payoff vectors have constant payoff efficiency. At the same time, in this measure zero subset, the payoff efficiency moral rule does not have any explanatory power at all since all payoff vectors have the same payoff efficiency. And, there is a voluminous literature that reveals economists' interest in behavior in such settings. Thus we apply our method to choice settings

[8] This result also may depend on the distribution from which random draws are taken. We use a uniform distribution. However, it is not inconceivable that the results will change if one takes some other distribution, say truncated normal.

of this kind and again ask whether the A-norm generally corresponds to another widely studied moral rule: inequality aversion.

Figure 5 shows the percentage of cases (out of 500,000 randomly generated sets of payoff vectors from a simplex defined by the condition $\sum_{i\in N} u_i(c) = 1$ and by the condition that all payoffs are non-negative) in which the choice-set-dependent and choice-set-independent inequality aversion models favor the same allocation as the A-norm. A first interesting observation is that both models generally do better than chance in matching the injunctive norm. However, the choice-set-dependent model does much better than the choice-set-independent model in settings with only a few consequences. When the choice set is small people may choose to use relatively complex choice-set-dependent norms. When the choice set is large the simpler choice-set-independent model performs about as well as the more complex model. Thus, people may prefer the simpler moral rule in this case, since it approximates the injunctive norm derived from the A-model almost as well as the more complex choice-set-dependent rule and is also easier to compute.

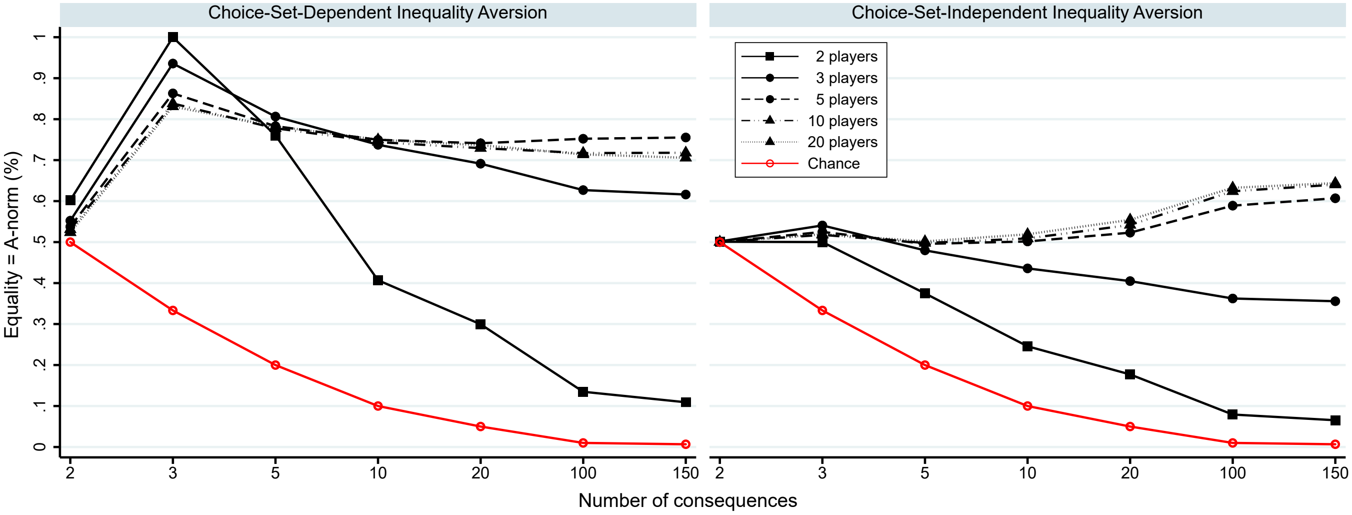

Figure 5: The percentage of A-norms that coincide with predictions of Choice-Set-Dependent and Choice-Set-Independent Inequality Aversion for random sets of payoff vectors with constant payoff efficiency. 500,000 sets for each case.

Conjecture 3. *In choice sets with fixed or little varying payoff efficiency where inequality becomes the main driving force of normative disagreement, we should expect that in small choice sets the A-model or complex choice-set-dependent heuristics are used. Knowing that this may be so can help discover such heuristics and thus better understand the choice in these conditions. When the choice set and the set of players are large a simple choice-set-independent heuristic should explain significant proportion of behavior.*

Thus far, we have deliberately focused on choice problems for clarity of exposition and to emphasize that the normative ranking of outcomes generated by our model depends only on the payoffs at each outcome (and the social weights). Nevertheless, it is useful to explore what injunctive norms look like in various widely studied strategic interactions (i.e., games), while cautioning that the model does not directly predict that people will choose the normatively most appealing outcome. It is worth reiterating that the injunctive norms (and moral rules) described herein should be thought of as inputs to a player's utility-maximization problem; it tells players what they *ought* to do, but in models of norm-dependent preferences, this is still traded off against what they *want* to do (i.e., against own-utility maximization). As we show in Kimbrough and Vostroknutov (2020), strategic considerations and heterogeneity in individual preferences for following norms come together to predict the actual decisions that people will make. With this caveat in mind, we next ask what kinds of moral rules we would expect to be salient on the basis of the A-model in social dilemmas, coordination games, and tournaments, under the assumption that social weights are equal for all players. For obvious reasons, heterogeneity in social weights may complicate matters, but we will note a few cases in which a little hierarchy would resolve an otherwise tricky normative ambiguity.

5.4 Moral Rules for Social Dilemmas

An obvious moral rule for social dilemmas is that players ought to cooperate, and indeed this is frequently, but not always, consistent with the injunctive norm as derived from the choice-set-dependent A-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 so high that the injunctive norm favors the outcome cooperate/defect (or defect/cooperate); this is because efficiency considerations can dominate in such extreme cases. Nevertheless, in most social dilemmas the moral rule "players ought to cooperate" will neatly capture the injunctive norm.

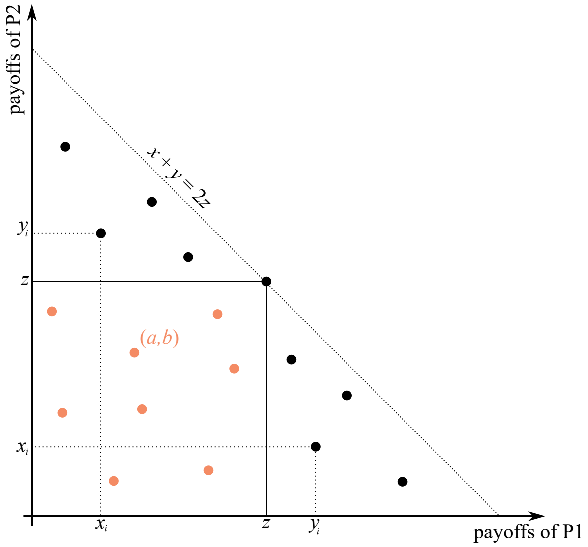

Figure 6: Symmetric two-player social dilemma.

We provide formal analysis that generalizes this point and consider a class of symmetric two-player social dilemmas in which the A-norm corresponds to the most efficient outcome even though for both players there are other consequences that bring them higher utility. Consider the set of payoff allocations for two players shown in Figure 6. Point (z, z) represents the most efficient symmetric allocation, or 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 the A-norm.

Proposition 8. *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 number of other points (a_i, b_i) with $a_i \leq z$ and $b_i \leq z$. Then (z, z) is the A-norm (has the least A-dissatisfaction).*

Proof. See Appendix A.

To provide additional intuition about what kind of behavior is implied by the injunctive norm in social dilemma games consider Figure 7 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 codes the aggregate dissatisfaction, with dark red being the outcome with the least dissatisfaction and dark blue the outcome with the most dissatisfaction.

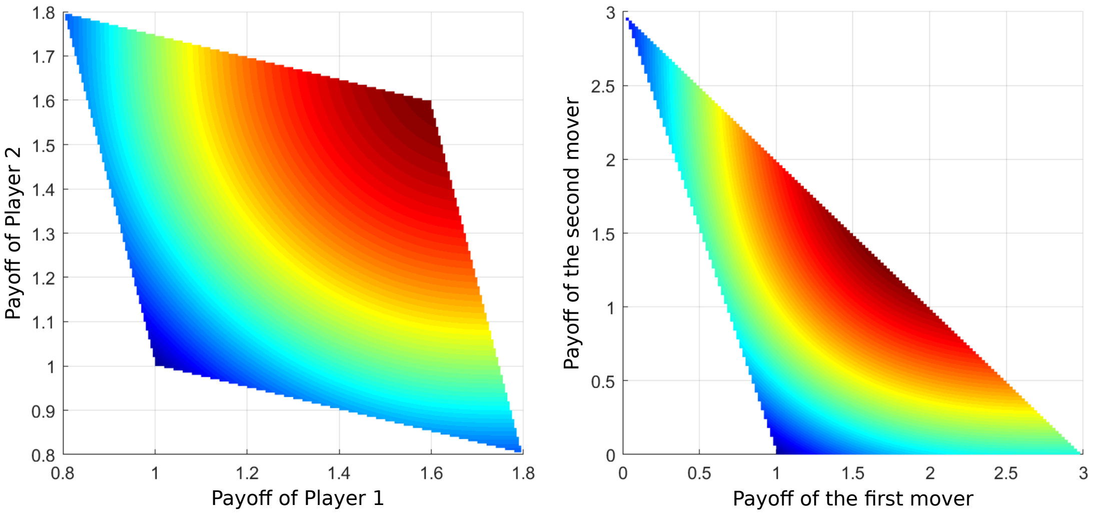

Figure 7: Left graph: the choice-set-dependent injunctive norm derived from the A-model in a 2-player Public Goods Game. Right graph: the choice-set-dependent injunctive norm derived from the A-model in a Trust Game. Dark red shows the most appropriate consequence and deep blue shows the least appropriate one.

The most appropriate consequence in the PG is for both players to contribute the whole endowment, which follows from Proposition 8, and the most inappropriate consequence is to contribute nothing. The normatively best outcome satisfies moral rules about maximizing efficiency and equality.

In the TG 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 8. It is also worth noting that for any choice of the first mover, the A-norm 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.

5.5 Moral Rules for Coordination Games

In coordination games, the ability of the injunctive norm derived from the A-model to generate a coherent moral rule depends on what other assumptions we make about payoffs. In coordination games with multiple Pareto-ranked equilibria (minimum effort games, stag hunts), the injunctive norm 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 injunctive norm provides little guidance. 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 in computing the norm (e.g., if it is one of the two players' birthday in the battle of the sexes).

5.6 Norm-Free Choice Settings

While injunctive norms apply to a broad array of choice settings, there are some settings in which the injunctive norm simply provides no guidance about what one ought to do. That is, there exist environments in which the normative evaluation implied by the A-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* and consider the following definition.

Definition 1. *Call an environment $\langle N, C, u\rangle$ **norm-free** if $D(c) \equiv d$, where d is some constant.*

We do not (yet) have a characterization of the set of norm-free environments in terms of payoffs, and in fact, we suspect that there are no intuitively interpretable constraints that define them. Any environment with two consequences and constant efficiency is norm-free, as well as, for example, the environment with two players defined by $C = \{a, b, c\}$ and $u(a) = (0, 3)$, $u(b) = (3, 0)$, $u(c) = (1, 1)$. Nevertheless, there is an important class of norm-free environments that we would like to describe.

These are the environments that have a structure reminiscent of a tournament.

Definition 2. *Call $\langle N, C, u\rangle$ 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 $c : N \to \{x_1, x_2, ..., x_N\}$ and $u(c) = (c(i))_{i \in N}$.*

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. An important property of tournaments is that they are norm-free. We prove this result in a proposition.

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

Proof. See Appendix A.

Thus, built into our model of norms that account for the dissatisfaction of all interested parties, is the existence of situations in which norms have no bite. Under our model of norm-dependent utility (Kimbrough and Vostroknutov, 2020), players in such settings are expected to behave as self-interested utility maximizers, which makes perfect sense in case of tournaments.

6 Discussion

We provide an axiomatic model of injunctive norms based on plausible psychological assumptions. We assume that individuals' normative judgments result from considering the entire set of feasible outcomes and their prospective dissatisfaction with each outcome as compared to all of the other feasible outcomes. To turn individual rankings of dissatisfaction into an injunctive norm, we further assume that people employ empathy, in the sense that they know and consider the fact that all other people feel dissatisfaction in an analogous manner. We thus assume that injunctive norms emerge that rank outcomes according to the aggregate dissatisfaction that they would produce, with the most appropriate outcome in the feasible set being the one that minimizes aggregate dissatisfaction of interested parties.

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, 2020), we show that a functional form derived from this model has substantial explanatory power in experiments and in particular can account for many observations that cannot be readily explained by standard theory or in models of outcome-based social preferences. However, in this paper we provide a potential explanation for why social preference models have normative appeal (and why there are many cases in which they seem to fit the data well).

We introduce a second set of axioms in which dissatisfaction is computed not relative to all other feasible outcomes, but rather, relative to some ideal as specified by a moral rule: for example, maximize efficiency or minimize inequality. We show how such ideals are able to closely approximate social preference models, and then we ask, for a variety of possible choice settings: how often does the choice-set-dependent injunctive norm derived from the first set of axioms correspond to the ranking implied by the second set of axioms under a given moral rule? This allows us to identify choice settings for which a particular moral rule is likely to be salient, precisely because it accurately reflects moral intuitions as captured in the injunctive norm.

Thus the model provides a meta-theory of moral rules, as approximations of radically choice-set-dependent injunctive norms. We argue that the complexity of computing injunctive norms and the difficulty of summarizing their implications renders such approximations desirable. We then provide plausible moral rules for a variety of widely studied choice settings.

Finally, we think it is worthwhile to highlight a sense in which this work bridges a gap between cooperative and non-cooperative game theory. In our model, norms are constructed on payoff sets, and not on games with actions. This mirrors the typical approach in cooperative game theory, in which the modeler asks: given some feasible set of allocations, on which one(s) would people agree? Our axiomatization is thus essentially a piece of cooperative game theory. Using the same objects as cooperative game theorists (the set of allocations), we create a mapping onto that same set that identifies the elements that are best according to a predefined set of principles (in our case, dissatisfaction-minimization). However, rather than stop there, we propose to use this analysis as an input to the analysis of non-cooperative games (Kimbrough and Vostroknutov, 2020). Thus our 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.

References

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

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.

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.

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. (2020). A theory of injunctive norms. mimeo, Chapman University and Maastricht University.

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

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

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

MacFarquhar, L. (2016). *Strangers Drowning: Impossible Idealism, Drastic Choices, and the Urge to Help.* Penguin.

Appendix (for online publication)

A Proofs

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

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

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

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

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

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

as desired. ▲

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

Proof of Proposition 3. $(1 \Rightarrow 2)$. For all $t_1, ..., t_N \in \mathbb{R}_+$ A2 implies $G(t_1, ..., t_N) = G(0, ..., 0) + \sum_{i \in N} \omega_i t_i$. By A1, $G(t_1, ..., t_N) = \sum_{i \in N} \omega_i t_i$. Thus, since D_i satisfy A5-A6 and d_i satisfy A1-A4, we have

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

as desired. ▲

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

Proof of Proposition 4. $(1 \Rightarrow 2)$. By B2, $f(0,0) = 0$ and by B1, $f_i(t,r) = f_i(0, r-t)$. So if $r-t \geq 0$ then $f_i(t,r) = f_i(0,1)(r-t) = r-t$. The last equality is by B4. When $r-t < 0$, by B2 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. ▲

$(2 \Rightarrow 1)$. B1-B4 are trivial. ■

Proof of Proposition 8. 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.

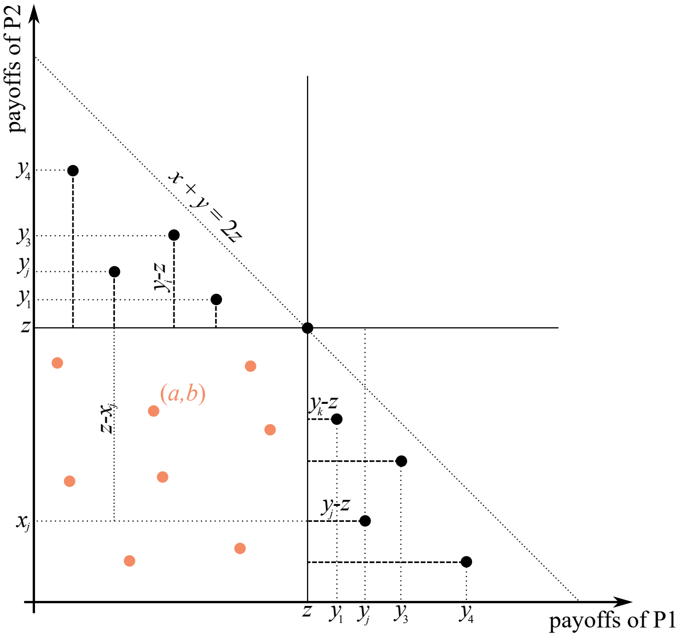

Figure 8: Illustration of the dissatisfaction calculations.

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 $z \leq y_1 \leq y_2 \leq \ldots \leq y_n$. Therefore, the dissatisfaction of any point (u_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 1 of Kimbrough and Vostroknutov (2020)). This makes (z, z) the A-norm. ■

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

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

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