A Meta-Theory of Moral Rules[*]

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

November 10, 2023

Abstract

Social interaction is, in part, governed by moral considerations. Moral judgments are complex because deciding what is right - in the absence of a clear rule - requires us to understand how others feel and how they would feel in alternative futures. Such judgments are all the more complex because they are heavily context-dependent: they are sensitive to which possible future scenarios are considered. One way people deal with moral complexity is by simplifying judgments via reference to some rule (e.g. splitting the bill equally, queuing, drawing lots). Such *moral rules* must generally cohere with underlying moral judgments if they are to be accepted as a substitute. In this paper, we formalize this idea by comparing two axiomatic models of moral judgment: one derives the notion of what is right from comparisons to some externally given, abstract ideal, and the other derives it from agents' preferences over the set of possible outcomes in the context. We show that the latter model is flexible enough to capture observed context-dependence in human behavior, but is computationally complex. The former model, because it abstracts from context to some degree, is less flexible, but computationally simpler and more readily articulated and codified. This explains the broad appeal of moral rules. It also suggests a method of predicting which moral rules might arise: by computing the context-specific moral judgment for a wide variety of related scenarios and articulating a rule that approximates those judgments.

JEL classifications: C02, C63, D63, D91
Keywords: injunctive norms, moral rules, axiomatic preferences, computational complexity.

[*] We would like to thank Valerio Capraro, David Rojo-Arjona, Matthias Stefan, and a seminar audiences at Maastricht University and the University of Konstanz for helpful comments. 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

Rules—instructions that tell us how to act in particular contexts—permeate our daily lives: we brush our teeth twice daily to keep them healthy; we stop at a stop sign when driving to avoid collisions with other drivers; etc. From an economist's perspective, it is not hard to imagine why such rules may emerge: following a rule simplifies individual decision-making in routine situations, which reduces computation and other transactions costs and helps us avoid undesirable individual outcomes. There may be an optimal time to brush one's teeth, based on diet, sleep, and other factors, but few people approach the practice of dental hygiene as an optimization problem. Following rules is also essential to creating and maintaining social order. Rules often play pivotal roles in sustaining institutions and allow people to work together in a more efficient and coherent way than would be possible if new decisions had to be made in each specific case. Introducing new rules or changing the existing ones can have a significant economic impact, thus making such rules not only an important subject of study, but also a potentially valuable policy lever.

Rules that guide private behavior can vary greatly from person to person. Some people, for example, may follow a complex sequence of procedures before going to sleep (e.g., applying different lotions), whereas others might have very simple bedtime rules (e.g., just brush the teeth). In private situations, people adopt individual rules and heuristics that make their lives simpler from the perspective of their own, individual preferences. Something similar can be said about social behavior in small groups of acquaintances (e.g., friends) where idiosyncratic rules may emerge from shared individual preferences; for example, friends might decide to always wear black or always meet for a 4pm coffee on Tuesdays. This suggests that there may be not that much global structure to rules of behavior of this kind. It might not be easy, if possible, to find regularities or predict on a population level how and why such individual rules emerge.

The situation with general rules of social behavior—rules that are agreed upon by most members of a society—is very different. Such rules cannot be too idiosyncratic; they need to be simple enough and expressed in a way that any member of the society can understand. Moreover, because these rules often carry with them an injunction (and often a threat of some punishment for violation), it is also essential that such rules generally comport with individuals' own views on what is right or socially appropriate. If a rule is too incompatible with individual judgments it will not be accepted. We call these rules that carry an injunction *moral rules*.[1]

Importantly, these observations make moral rules amenable to scientific inquiry. The fact that not only the content of a moral rule, but also the severity of its violation must be (more or less) agreed upon by most members of a group for the rule to have force suggests that moral rules will have structure. This structure will derive from the commonalities present among the individual normative judgments of

[1] This distinguishes them from "mere conventions", which lack the same moral force. For example, a rule of sharing is different from a rule about which side of the road we drive on, because the former reflects a shared normative commitment that shapes the content of the rule, while the only normative commitment undergirding the latter is that there ought to be *some rule*, though any will do.

group members. This means that we should expect many moral rules to reflect, in general, the kinds of social norms that emerge from balanced consideration of judgments of individual group members. This means that a theory of rules can be constructed on top of a theory of individual moral cognition.[2]

To illustrate the argument, we build on the theory developed by Kimbrough and Vostroknutov (2023), who have shown one way to create a model of individual-level, context-dependent moral cognition and aggregate it into shared *injunctive norms* that balance the interests of all agents involved in a particular game or allocation-problem.[3] Common knowledge of what each person desires and of how counterfactual outcomes compare to one another from each person's point of view can be combined to find a *dissatisfaction minimizing* outcome that the authors argue will be seen as normative. Here we point out that, because of its radical context-dependence, such a model requires both large amounts of information and complex computation to find an injunctive norm—deficiencies that could be reduced if it were possible to summarize these aggregated, context-dependent norms via a simple rule.

Thus, in this paper, we build on the method of modeling normative decision-making introduced by Kimbrough and Vostroknutov (2023), and we show how to extend it to agents whose judgments of each outcome depend on the correspondence between that outcome and some abstract moral rule (in a given context). In their model, agents experience dissatisfaction when outcomes give them less than they could receive at counterfactual outcomes; in this model, agents experience dissatisfaction when outcomes deviate from the outcome prescribed by a moral rule. Thus, we introduce an axiomatic framework for constructing such moral rules.

The axioms that we propose do not constrain the content of moral rules—in principle, moral rules can be arbitrary, but in practice, we observe regularity in the kinds of moral rules that people coordinate on. As noted above, we attribute this to the observations that moral rules should be simple and, for them to be effective, they cannot deviate too much from individual normative judgments. That is why, in the next step we suggest that the *need* for a moral rule in a class of contexts and its specific *content* are both determined by the properties of the injunctive norms, as defined in Kimbrough and Vostroknutov (2023), that arise in these contexts. This leads to two intuitive conclusions: (1) a moral rule is more necessary, the more complex the computations required to identify the context-dependent normative ranking of outcomes, and (2) a moral rule's content must approximate the pattern of judgments that result from context-dependent individual moral cognition, or else it will not be viewed as acceptable.

The idea that we often rely on simple moral rules to reduce the costs of complexity is easy to see in the way checks are often divided among colleagues or friends at a restaurant. Rather then pay the costs of determining precisely who ordered what and how much they owe, we often instead agree to simply divide the check by N, and we groan if someone wants to divide things differently. This rule sometimes leads

[2] Thus, we are advocating a theory of morality in which—although we sometimes behave as if the notions of right and wrong can be derived from some rule—the kinds of rules we articulate are derived from (or constrained by) a more fundamental moral psychology that governs our sense of what is right and wrong. See, e.g., Wilson (2020) for a discussion of this idea in the context of property rights.

[3] We have reasons for thinking that this model is especially useful, but the point is simply that we need some model of moral cognition to juxtapose to a model of moral rules.

to dissatisfaction—someone may know that they, who only ordered an entree, are paying the same as another person also ordered a salad and a dessert. If the rule imposes too much, then it might be protested, but for the most part, people accept these minor asymmetries to simplify the decision. Following a moral rule often means sacrificing context-dependent precision in doing what is right for a reduction in the costs of figuring out the right thing to do.

The cost advantage of moral rules is especially evident in moral teaching: children are often first taught explicit moral rules of behavior ("don't take", "don't hit", "share equally", "wait your turn") because these are simple to communicate and mitigate predictable interpersonal conflicts. However, as children age, we begin to teach them why such rules exist: "imagine how *you* would feel if someone did that to you," "think about what would happen instead if we didn't follow the rule", etc. Applying empathy and counterfactual reasoning to determine what is right is challenging, and we learn to do it rather painstakingly, over time. In doing so, we come to note certain incongruities between the actions required by moral rules and the actions that would satisfy our moral intuitions. Such mismatch—if sufficiently glaring—generates protest and constrains the set of possible moral rules to those which do not deviate too much from our moral judgments.

We formulate these ideas as a *meta-theory of moral rules.* In Section 2, we start with the axioms that describe our general definition of a moral rule: a moral rule is a specific preference relation over a set of feasible allocations (aka a context). The fact that we allow moral rules to depend on the context provides them with a degree of context-dependence. This idea comes from our conviction that moral rules (and rules in general) must be context-dependent, given that life presents us with many possible contexts, and rules by definition cannot apply in the same way to all of them. For example, a stop sign makes sense only in the context of driving, and does not make sense when you see it standing in the middle of a forest. We provide examples of how well-known concepts like Pareto efficiency, maximin, preference for efficiency, and various types of inequality aversion can be described as moral rules. Then, we discuss strengths and weaknesses of moral rules and show with a little experiment that the degree of context-dependence inherent in otherwise appealing moral rules is not enough for them to capture moral intuitions in all contexts.

In Section 3, we compare our slightly-context-dependent moral rules with the radically-context-dependent injunctive norms of Kimbrough and Vostroknutov (2023), further KV. We prove a general result showing that moral rules are not equivalent to injunctive norms in the sense that no moral rule can capture the radical context-dependence of injunctive norms as defined by KV. In other words, following moral rules can increase dissatisfaction from the point of view of a context-dependent model of individual moral cognition. We argue that this helps explain the observation that people adopt different moral rules in different contexts.

In Section 4, we introduce a measure of the computational complexity of applying moral rules and injunctive norms. The measure suggests that the complexity of moral rules grows linearly with the number of feasible allocations in a context, while the complexity of injunctive norms grows quadratically. This shows that the cost of moral rules in terms of their normative imprecision may be compensated by the relative ease of computing and communicating what ought to be done according to the rule. We argue that this can help explain why moral rules remain an integral part of moral reasoning: people may be willing to sacrifice context-dependent precision in doing what is right for a reduction in the costs of figuring out the right thing to do.

In Section 5, we show how plausible moral rules can be derived from injunctive norms. In some specific classes of contexts the rules can be derived analytically from our model of injunctive norms (see KV). Such rules are then the precise representations of injunctive norms, which makes them very likely to appear. However, in most contexts such analytical derivations do not produce simple solutions that can be easily articulated as a moral rule. For such cases, we use simulations and show that simple rules like efficiency and maximin fit remarkably well on average in classes of general random contexts (they identify the same outcome as maximally appropriate that is identified by injunctive norms). Then we show that inequality aversion rules fare well in classes of contexts with fixed efficiency (like Dictator games). The catalog we produce is far from exhaustive, but we argue that these kinds of simulations provide a method to determine which moral rules are likely to emerge in which specific classes of contexts (for a given model of moral psychology).

In Section 6, we summarize our ideas as a meta-theory of moral rules, which suggests that the reason moral rules exist is because computing morality can be costly and moral rules play the role of cheaper, but good enough, substitutes for precise injunctive norms. Finally, in Section 7 we discuss some examples and perspectives on future directions for research.

2 Dissatisfaction Functions for Moral Rules

As in KV, we begin with the assumption that normative judgments are individual and comparative. By the former we mean that people are able to perceive how normatively appropriate their own situation is, from the perspective of their *individual consumption*; the latter means that the normativity comes from the *comparison* of the current situation to some counterfactual. Then, we assume that, in a given situation, people are capable of perceiving the individual normative judgements of others through empathy. A moral rule emerges through *aggregation* of individual normative judgements by each person into a common perception of social appropriateness.

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.* 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 context 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* context-dependent moral rule—which defines the normatively best element in any choice set—with a dissatisfaction function.

Consider a large set $\mathcal{C}$ of all possible consequences (e.g., outcomes of games), N players, and a value function $u : \mathcal{C} \to \mathbb{R}^N$, which for each consequence defines a vector of consumption values associated with the payoffs of a game or choice problem. 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 $C \subset \mathcal{C}$. Together, C and the collection of associated payoff vectors $u[C]$ can be thought of as the set of all feasible allocations in the context of some 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 normatively appealing, i.e. they yield zero dissatisfaction, because they reflect the ideal. In addition, there is a function $u^*(x \mid C^*)$ that, for each allocation $x \in C$, defines an element of C^* as a "reference point" for $x \in C$. The function should be defined for each vector $x \in C$ and all subsets of $\mathbb{R}^N$ that are equal to $u[C^*]$ for some C.[4] 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: $u^*(x \mid C^*) = x$ whenever $x \in C^*$, 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, we consider functions $f_i : \mathbb{R}^2 \to \mathbb{R}_+$ that define *dissatisfactions* for each player $i \in N$. Namely, $f_i(u_i(x), u_i^*(x \mid C^*))$ stands for the dissatisfaction that player i feels when her consumption value is $u_i(x)$ and the value that she would receive in the ideal situation is $u_i^*(x \mid C^*)$. 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 context-dependent model of normative judgment. For notational convenience and to be consistent with radically context-dependent axioms defined in KV, we will denote individual normative evaluations by $F_i(x \mid C) = f_i(u_i(x), u_i^*(x \mid C^*))$, which we call *total dissatisfaction function* of player i.

In the next step, we define an *aggregate dissatisfaction function* $F(x \mid C)$ that creates a composite of the dissatisfactions 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 functions f_i, total dissatisfaction functions F_i, and the aggregate dissatisfaction function

[4] Specifically, there is no need to define $u^*(x \mid C^*)$ for all subsets 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 context-independent dissatisfaction with only one C^* for all C in Section 2.1. There we need to define u^* for only one set C^*.

F. We start with the axioms that define the connection between dissatisfaction f_i and the consumption value of payoffs received by player i.

R1 $\forall i \in N\ \forall C \subset \mathcal{C}\ \forall x, y \in C\ \forall a \in \mathbb{R}\quad f_i(u_i(x) + a, u_i(y) + a) = f_i(u_i(x), u_i(y))$.

R1 states that adding a constant to both the consumption value of some consequence x and the consumption value of the ideal y, that are being compared, does not change the dissatisfaction. So, if player i gains or loses the same amount of consumption value in all consequences (plus ideal), then her dissatisfaction is unaffected. R1 ensures that f_i can be expressed as a function of the difference between the agent's value at x and at y.

R2 $\forall i \in N\ \forall C \subset \mathcal{C}\ \forall x, y \in C\ \forall \alpha \in \mathbb{R}$ with $\alpha > 0\quad f_i(\alpha u_i(x), \alpha u_i(y)) = \alpha f_i(u_i(x), u_i(y))$.

R2 states that if all payoffs are multiplied by a positive constant, then the dissatisfactions are also multiplied by the same constant. This ensures that dissatisfactions are proportional to consumption value in a linear way, thus connecting the two concepts. We could have assumed some non-linear, say concave, relationship between dissatisfaction and value. However, we already allow consumption value to be a non-linear function of payoffs. R2 reflects an idea that all non-constant marginal effects of payoffs are already encoded in functions u_i.

R3 $\forall i \in N\ \exists \beta_i \in (0, 1]$ such that $\forall t \geq 0$ we have $f_i(0, -t) = \beta_i f_i(0, t)$.

R3 states that there might be an asymmetry in how dissatisfaction is perceived when the ideal outcome yields consumption value above, versus below, that generated by x. Specifically, if player i receives 0 when the ideal yields $-t$, then her dissatisfaction could be lower than the dissatisfaction she gets when the ideal yields 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 consumption value is greater than what she would receive at the normatively ideal outcome. However, then it would be possible that there are consequences for which aggregate dissatisfaction is zero (all players get higher values than they would at the ideal), 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 outcomes that have zero dissatisfaction but that are not included in C^*.

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 and there exists another consequence that gives them 1. This assumption amounts to the claim that the dissatisfaction from the same change in consumption value makes all players dissatisfied in the same way. This is how we operationalize the idea that shared normative evaluations are built on empathy. Empathy is the mechanism of interpersonal comparison.[5]

The following proposition provides the representation.

Proposition 1. The following two statements are equivalent:

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

Proof. See Appendix A.

The next axiom defines the total 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 value in x (see Section 2.1). This highlights the sense in which ideals involve abstraction; the ideal may not be among the feasible consequences in C.

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

R5 says that 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, R5 asserts that for any two sets C_1 and C_2 that have $u[C_1^*] = u[C_2^*]$, or which have the same ideal, i's personal dissatisfaction of all elements $x \in C_1 \cap C_2$ is the same. As we see below in Section 3.2, this feature of total dissatisfaction for moral rules lies at the root of their limited context-dependence and makes moral rules less flexible than radically context-dependent injunctive norms of KV. 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 $F_i(x \,|\, C_1) = F_i(x \,|\, C_2)$ for all $x \in C_1 \cap C_2$ and all $i \in N$.*

Proof. By the property of u^*.

Next, we aggregate the dissatisfactions across players and define the aggregate dissatisfaction function 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 $F(x \,|\, C)$ is a function of $F_i(x \,|\, C)$ for all $i \in N$. Specifically, we assume

[5] It is interesting to think about the consequences of imperfect empathy, which could be a source of normative conflict, if people incorrectly assess how others value various outcomes. We leave this for future work.

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 $\omega_i \in \mathbb{R}$. These weights determine how much each player's dissatisfaction counts in the computation of aggregate dissatisfaction, and they can represent power, social status, in/outgroup relationships, kinship, or their combination (e.g. negative weights could "legitimize" moral rules that prescribe outright hostility towards others).

R7 $\forall i \in N \; \forall F_1, ..., F_N \in \mathbb{R}_+ \; \forall a_i \geq -F_i \quad E(F_i + a_i; F_{-i}) = E(F_i; F_{-i}) + \omega_i a_i$.

The notation $E(F_i; F_{-i})$ 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 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
$$F(x \mid C) = \sum_{i=1}^{N} \omega_i F_i(x \mid C) = \sum_{i=1}^{N} \omega_i (\max\{u_i^*(x|C^*) - u_i(x), 0\} + \beta_i \max\{u_i(x) - u_i^*(x|C^*), 0\}),$$
where $\beta_i \in (0, 1]$ *are some 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

$$w_i(x \mid C) = u_i(x) - \phi_i F(x \mid C).$$

The main point is simple: player i trades-off personal consumption value $u_i(x)$ and conformity to moral rules, $-F(x \mid C)$, defined as the *negative* of aggregate dissatisfactions (since they ought to be minimized). Player i makes this tradeoff according to her idiosyncratic propensity to follow norms, $\phi_i \geq 0$. In KV, we call $\eta(x \mid C) = -F(x \mid C)$ a normative valence of x in C and η a norm function. 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 2.2 below.

2.1 Examples of 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 $u^*(x \mid C^*)$ can be defined as an element of C^* that is the closest to x 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 $y \notin C^*$. $u^*(x \mid C^*)$ can be defined as an element of C^* that is the closest to x 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 $y \notin C^*$. $u^*(x \mid C^*)$ can be defined as an element of C^* that is the closest to x in Euclidean metric.

Cooperative Solution Concepts Broadly speaking, axiomatic models drawn from cooperative game theory can be (re)interpreted as models of moral reasoning and can then be used to provide a foundation for moral rules. 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.

Context-Dependent Inequality Aversion Take any C and compute the average payoff that each player gets in all consequences: $a_i^* = \sum_{x \in C} u_i(x)/|C|$. 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^*.[6]

[6] A similar but less payoff-sensitive principle of inequality aversion might take into account only the highest and the lowest payoffs that players receive in C. For each player compute $\underline{m}_i^* = \min_{x \in C} u_i(x)$ and $\overline{m}_i^* = \max_{x \in C} u_i(x)$. 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 context-dependent: adding any consequences that do not change $\underline{m}_i^*$ and $\overline{m}_i^*$ does not change the ideal and thus does not affect the dissatisfaction of other elements in C.

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, $F_i(x \,|\, C) = F_i(x) = f_i(u_i(x), u^*(x \,|\, C^*))$. 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 context-independent dissatisfaction function.

Context-Independent Inequality Aversion Let $C^* = \{x \in \mathcal{C} \,|\, \forall i, j \in N \;\; u_i(x) = u_j(x)\}$. This is a line 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 $u_i^*(x \,|\, C^*) = (a, ..., a) \in C^*$ be the average payoff (identical for all players).

2.2 Strengths and Weaknesses of Moral Rules

To build intuition, we illustrate the strengths and weaknesses of 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 offer 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 context-dependent inequality 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, context-independent inequality aversion).

From this example, it may seem that context-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 context-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 context-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^*.

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.[7] We

[7] 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 $k = 1..|C|$.

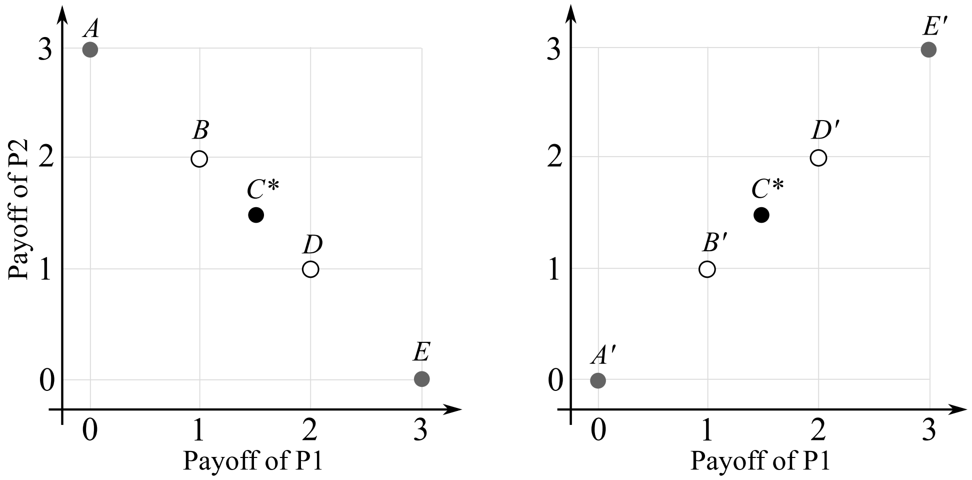


Figure 1: **Left.** Context-dependent inequality aversion in the context of the Dictator game with allocations A, B, D, E (the ideal is C^*). White circles denote the most appropriate allocations with respect to C^*. **Right.** Same but in the context A', B', D', E' obtained from the Dictator game by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

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 context-dependent inequality aversion—C^* is the same in the EG as it was in the Dictator game. 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 context-dependent inequality aversion (marked with white circles).

To test this intuition, we conducted an online experiment on Prolific with 100 people from the UK. 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.[8]

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 context-

[8] We chose one of the four 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 B. This experiment was conducted with approval from the Maastricht intercity ethics committee under their common IRB agreement with the BEELab at Maastricht University ($ERCIC\_379\_28\_09\_2022$).

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, context-dependent inequality aversion—the only rule capable of accounting for DG choices in List (2007)—is unable to account for EG choices, as efficiency considerations take over.

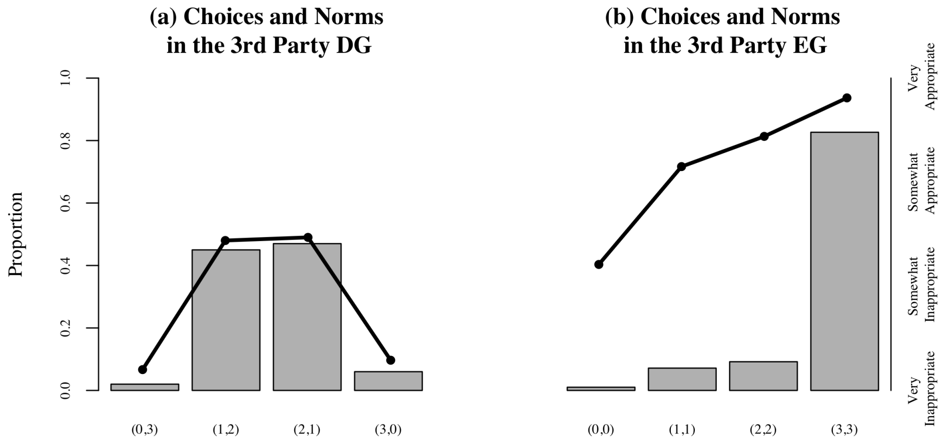


Figure 2: Choices and elicited norms in the 3rd party DG and EG shown in Figure 1.

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 can be easily expressed and even 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 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 compare a model of endogenously arising, radically context-dependent injunctive norms of KV with our treatment of moral rules.

3 Comparison of Moral Rules and Injunctive Norms

3.1 Injunctive Norms

The example above is an instance of the general phenomenon that a given moral rule, although context-dependent by construction, might not produce reasonable results in *all* environments. Extensive experimental evidence also suggests that different moral rules (e.g., inequality aversion, payoff efficiency, maximin, etc.) seem to be used in different environments, and that one person can often switch from one moral rule to another.[9] Given these problems with moral rules, KV propose another set of axioms that produce aggregate dissatisfaction functions that account for such radical context-dependence.

Specifically, they assume that player i at consequence $x \in C$ is personally dissatisfied when there are other consequences that give her higher consumption value than $u_i(x)$, and norms emerge that balance these self-interest motivated dissatisfactions across interested parties. This construction makes the ideal more dependent on the set of feasible outcomes than the moral rules discussed above. Whenever two contexts differ by a single outcome that yields a higher payoff than at least one alternative outcome for some individual, that individual will re-evaluate all the less preferred alternatives as less appealing, and the norm will adjust accordingly. KV provide a representation result for this class of injunctive norm functions (for this case, aggregated dissatisfaction function is called D instead of F):

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

Here, total dissatisfaction $D_i(x\,|\,C)$ of i at x is the sum of her dissatisfactions due to *all* other consequences in C that give her higher consumption value. The function $-D(x|C)$ then enters the norm-dependent utility as the norm function generated by the injunctive norms derived from minimizing the sum of all individual dissatisfactions. KV show that this representation can account for observed behavior in a wide variety of experimental designs.

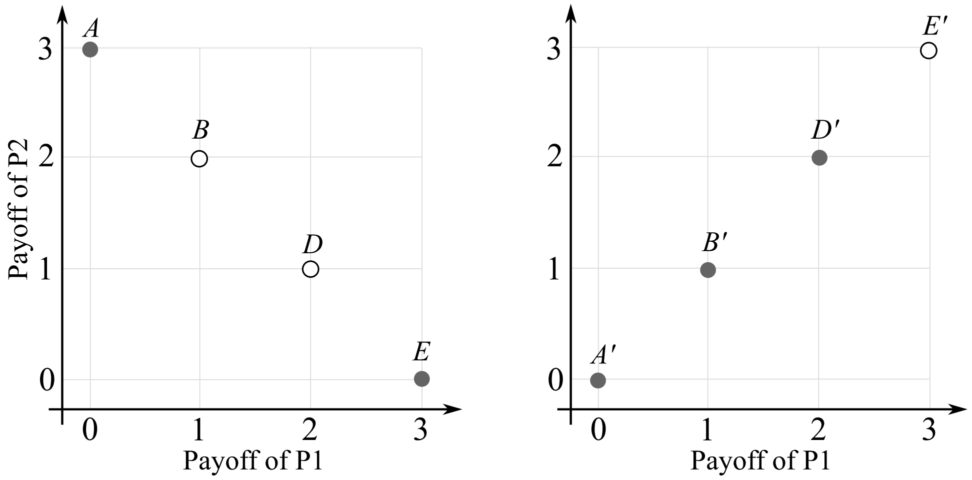


Figure 3: **Left.** Radical context-dependence in the Dictator game with allocations A, B, D, E. White circles denote the most appropriate allocations. **Right.** Same but in the context A', B', D', E'.

To illustrate, consider Figure 3 that shows the same environments as in Figure 1 but with the radically context-dependent representation of the ideal. The allocations B and D are still the most appropriate in the left panel (marked with white circles) but unlike in Figure 1 the more reasonable allocation E' is now

[9] See for example McCabe et al. (2003); Engelmann and Strobel (2004); List (2007); Bardsley (2008); Baader and Vostroknutov (2017); Galeotti et al. (2018).

the most appropriate in the right panel. Thus, we can see that the radical context-dependence is flexible enough to account for the apparent switches in moral rules that are detected in experiments.

3.2 Non-Equivalence of Moral Rules and Injunctive Norms

In this section, we generalize the idea above and show that norms generated from any moral rule are "less context-dependent," and as a result less flexible, than the radically context-dependent injunctive norms of KV. To do this, we need a notion of equivalence between them. It may seem that we could simply call a moral rule equivalent to the injunctive norm in some context C whenever they generate the same maximal elements (where aggregated dissatisfactions are minimized). However, this definition will not account for important information codified in the normative valences of other alternatives. This is because the entire ranking of outcomes is relevant for choice, given that interior solutions of the utility maximization problem with norm-dependent utility are sensitive to the relative normative valences of all the outcomes that are traded off against consumption. Thus, we compare the models by asking whether they can induce the same ranking of consequences in terms of aggregate dissatisfaction functions F and D.

Let $\succsim_C$ be defined as the preference relation that represents the aggregated dissatisfaction function D in some set C:

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

Similarly, for moral rules defined by the collection of all C and C^* together with a dissatisfaction function f, let $\succcurlyeq_C$ represent the dissatisfaction function F:

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

Let us say that a moral rule F is *equivalent* to injunctive norm D if $\succsim_C$ and $\succcurlyeq_C$ are the same for all $C \in \mathfrak{C}$. Our task now is to show that there exists no moral rule that is equivalent to D. To do that, let us try to construct a moral rule that generates aggregated dissatisfaction F that is as close to D as possible. Moral rules are defined through ideal elements C^*. So, for each C let us define C^* as the set of minimal elements of $\succsim_C$ (the preference relation induced by D) and choose any dissatisfaction function f.

The following example shows how no such moral rule can be equivalent to the injunctive norm function induced by D. 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 both $C_{(4,5)}$ and $C_{(5,4)}$, the injunctive norm, or the allocation with minimal dissatisfaction, is $(5,5)$. So, when we construct a moral rule as described above, for both contexts we set $C^* = \{(5,5)\}$. By Propositions 2 and 3, for any moral rule it is true that $F((3,6) \,|\, C_{(4,5)}) = F((3,6) \,|\, C_{(5,4)})$. The same also holds for allocation $(6,3)$. Thus, according to any moral rule 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 injunctive norms generated by D: here, the relative dissatisfactions of $(3,6)$ and $(6,3)$ change depending on y. The dissatisfaction of $(3,6)$ is smaller than that of $(6,3)$

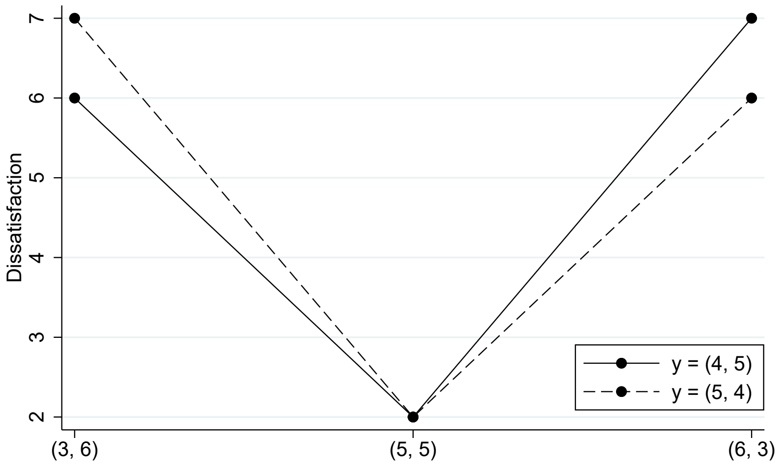


Figure 4: Dissatisfactions D of the common allocations in $C_y = \{(3,6);(5,5);(6,3);y\}$, where $y = (4,5)$ or $y = (5,4)$.

in $C_{(4,5)}$, but larger in $C_{(5,4)}$. Thus, no moral rule is equivalent to injunctive norms. We state this result as a proposition.[10]

Proposition 4. *There is no moral rule that is equivalent to the injunctive norm in the sense of generating equivalent preference orderings of aggregated dissatisfaction for all contexts.*

Proof. By example on Figure 4.

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

For more general moral rules 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, for the injunctive norms, 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 two models, which demonstrates that in the case when $C^* \subsetneq C$ equivalence cannot be achieved either.

Moreover, the injunctive norm that minimizes aggregate dissatisfaction D is never Pareto-dominated (see KV). Thus, if a moral rule has some C^* that contains Pareto-dominated consequences, as for example in the context-dependent inequality aversion, then this will always contradict the predictions of the injunctive norms model except for special cases like Dictator games with constant efficiency.

Finally, the discrepancies in induced preference orderings between injunctive norms and moral rules are not "rare" or measure zero: 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

[10] It may seem that we could construct an equivalent moral rule 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.

that arises from injunctive norms. Thus, we should expect systematic differences between injunctive norms and moral rules in terms of dissatisfaction rankings.

4 Why Moral Rules?

As shown above, the radical context-dependence of injunctive norms, in which the evaluation of each consequence depends on all other feasible consequences, outperforms moral rules that use only limited context-dependence based on ideals. If we hypothetically imagine that injunctive norms, as summarized by function $D(x|C)$, provide an account of peoples' normative intuitions, this poses a problem with understanding why moral rules even exist: moral rules are less flexible than injunctive norms and do not reflect moral intuitions in all contexts; when using moral rules, dissatisfaction $D(x|C)$ will not in general be minimized. What good, then, are moral rules? We argue that one might nevertheless prefer an inexact moral rule to a precise intuition of injunctive norms because of the latter's relative *complexity*.

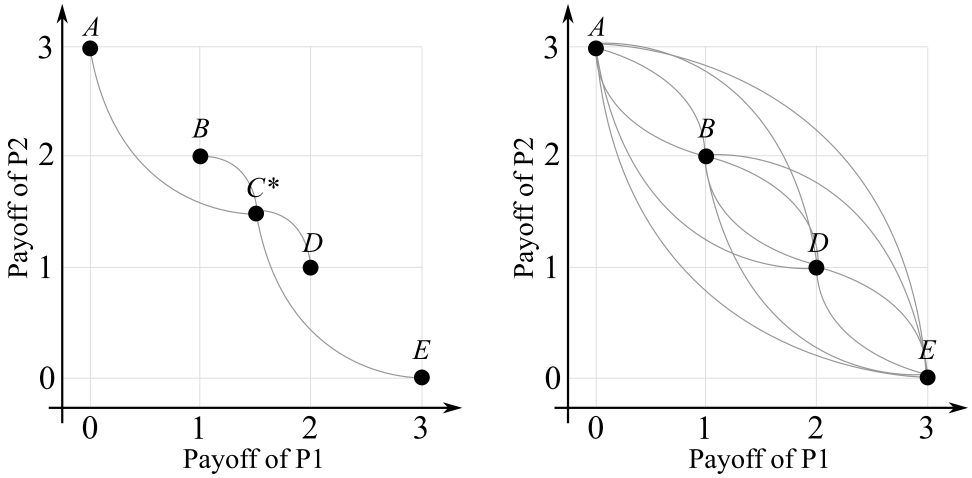


Figure 5: **Left.** Comparisons needed to compute a moral rule (per player). **Right.** Comparisons needed to compute an injunctive norm (per player).

Figure 5 shows with curvy lines the comparisons between allocations that must be made to compute dissatisfaction according to a moral rule (the left panel) and according to an injunctive norm (the right panel). For the moral rule, only 4 comparisons should be made: each allocation is compared to an ideal. For the injunctive norm however, we need to compare each allocation to every other, which brings the number of comparisons to 12.[11] Thus, already with a small context like this, the number of comparisons for the injunctive norm triples.

In general, suppose that there are N players (all with $\omega_i = \beta_i = 1$ for simplicity) and consider some set of consequences C that has k elements. Then, to compute a moral rule one needs to perform Nk payoff comparisons (and decide whether they are greater or equal to zero). So, the complexity of this problem can be denoted by $O(Nk)$. However, to compute the injunctive norm one needs to perform

[11] Each comparison consists of N "sub-comparisons," one for each player.

$Nk(k-1)$ payoff comparisons. We can say that the complexity of this problem is $O(Nk^2)$. So, for any fixed number of players, the moral rule can be computed in linear time, whereas the injunctive norm can be computed only in quadratic time. Therefore, as the number of consequences grows, so does the appeal of an abstract moral rule due to its reduced complexity.

This argument suggests that moral rules may be used in some situations instead of injunctive norms because they are easier to compute. Another reason why people may be attracted to moral rules is that rules can be articulated in a way that generalizes across choice settings (e.g., "one ought to maximize efficiency"), thus facilitating coordination and cooperation. The possibility to easily explain a moral rule, to write it down as a law or a religious principle allows people to communicate current normative arrangements to each other, thus making sure that more people understand and respect them. Conversely, radically-context-dependent injunctive norms do not have this property: people often cannot explain why they feel that some consequence is more appropriate than another in a specific context. This leads to a situation where people would need to recompute the intuition coming from injunctive norms each time anything changes, which can create large computation and transaction costs. The idea that moral rules are superior to injunctive norms when it gets to explaining normativity to others suggests that moral rules can also be very valuable for teaching normative ideas to children, which is necessary for maintaining proper functioning of institutions and societies overall.

5 Deriving Moral Rules from Injunctive Norms

5.1 Analytical Techniques

The previous section provided arguments 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 by injunctive norms. We thus ask, for various classes of choice settings, what are the commonly observed properties of injunctive norms generated in those settings that might be summarized as a moral rule?

KV describe various types of contexts in which injunctive norms can be summarized as moral rules. For example, they show analytically that regardless of the choice setting, the minimum of the aggregate dissatisfaction function $D(x|C)$ will be always Pareto optimal. This implies that injunctive norms are always consistent with Pareto moral rule ("choose an allocation that is not Pareto-dominated"). For the special case of choice settings with exactly two possible consequences, they show that injunctive norms imply payoff efficiency maximization. For the class of contexts where C is a convex N-polytope that satisfies a rather weak scarcity condition (satisfied, for example, by any asymmetric Public Goods or Trust games), they show that the minimum of $D(x|C)$ never gives maximal payoff to any one player, so that in the most appropriate outcome all players always have to compromise (Compromise Theorem in

KV). Thus, it is possible to derive moral rules from injunctive norms directly for specific classes of choice settings.

In general however, such analytical proofs may not produce any simple rule, and thus in this section, we use simulations to discover what moral rules may arise in different conditions. We append random payoff vectors to various numbers of consequences under different restrictions (e.g., constant sum) and ask how frequently the intuition from injunctive norms corresponds to various other plausible (or commonly expressed) moral rules: “maximize efficiency”, “minimize inequality”, etc. As it turns out, injunctive norms often cohere reasonably well with existing moral rules depending on the context.

5.2 Simulations

5.2.1 Random Payoff Vectors

From KV, we know that payoff efficient allocations minimize $D(x|C)$ for all choice sets with exactly two consequences. For larger sets of consequences, this is however not always true. Nonetheless, it is true quite frequently, enough so that it is plausibly reasonable to summarize this tendency of injunctive norms 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 minimum of the aggregate dissatisfaction function $D(x|C)$ (hereafter, D-norm) would also be chosen by the moral rule “choose the most efficient outcome”.[12]

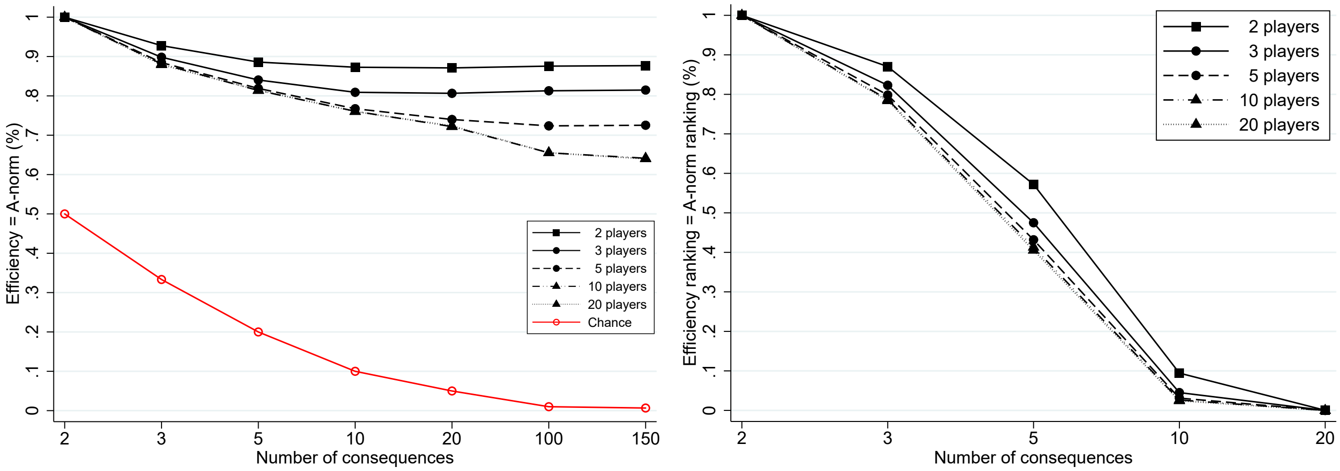


Figure 6: **Left.** The percentage of D-norms that are payoff efficient for random sets of consequences. 500,000 sets for each case. **Right.** The percentage of cases in which the injunctive norm function ranks outcomes according to their payoff efficiency.

The left panel of Figure 6 shows the results. With 3 players and a large number of consequences, around 80% of D-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 D-norms drops

[12] 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 $[0, 1]$.

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 injunctive norms is identical to the ranking according to payoff efficiency, or what percentage of the time the injunctive norm is equivalent to the moral rule in the sense of Section 3.2. The right panel of Figure 6 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 injunctive norms for small sets of random consequences and players. As these numbers grow, payoff efficiency becomes progressively worse at predicting the most normatively desirable outcome (D-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 D-norm in* 80*-*90% *of cases. As the set of consequences and the set of players grow, the payoff efficiency rule should be used less often.*

Next, we consider the maximin criterion as a moral rule for random payoff vectors ("choose the allocation that gives the highest payoff to the poorest player"). The left panel of Figure 7 shows the percentages of cases in which the D-norm coincides with the maximin ideal. Overall, the maximin rule corresponds to the D-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, another 40% follow the payoff efficiency criterion, which also corresponds to the D-norm quite often in such situations: for 3 players and 3 consequences the number of payoff efficient D-norms is 90% and the number of maximin D-norms is around 75% (see Figures 6 and 7).

KV argue that maximin is more likely to be the D-norm when players take the logarithm of payoffs to compute dissatisfactions. The reason is that the log 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 7 shows the percentages of cases in which the D-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 7—especially for the

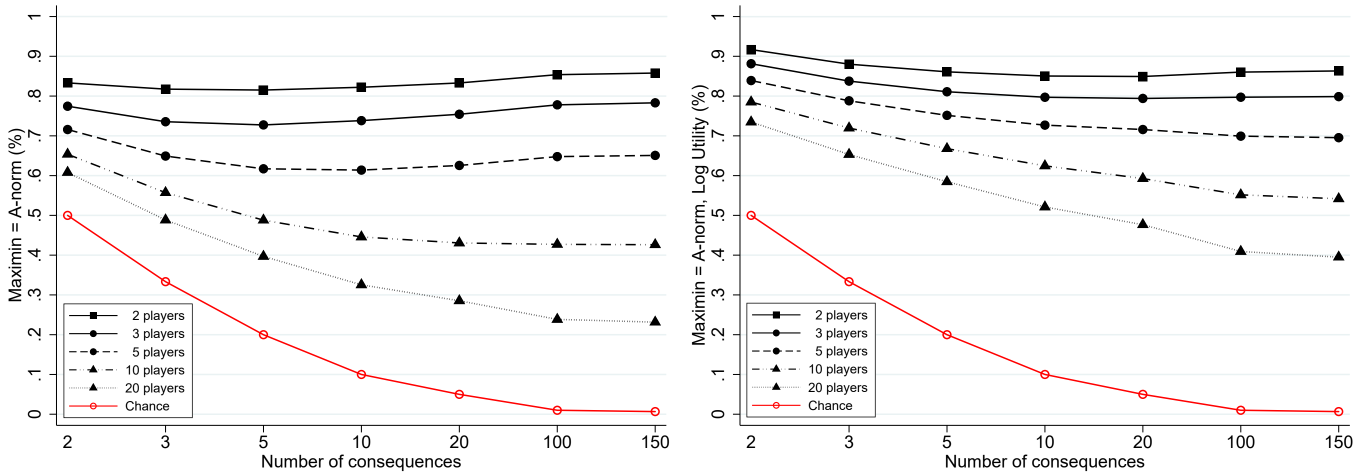


Figure 7: **Left.** The percentage of D-norms that are maximin for random sets of consequences. **Right.** The percentage of D-norms when consumption value is logarithmic in payoffs that are maximin for random sets of consequences. 500,000 sets for each case.

large sets of players and consequences. This provides support to the intuition in KV. 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 D-norm in* 70*-*90% *of cases under the assumption that people's value functions take the logarithm of payoffs. This also holds to a lesser degree even without the logarithm when consumption value is linear in payoffs. As the set of players grows, the maximin rule should be used less often. Given that both payoff efficiency and maximin fit the D-norm rather well when the set of players is small, we should expect to observe both rules used in these circumstances.*

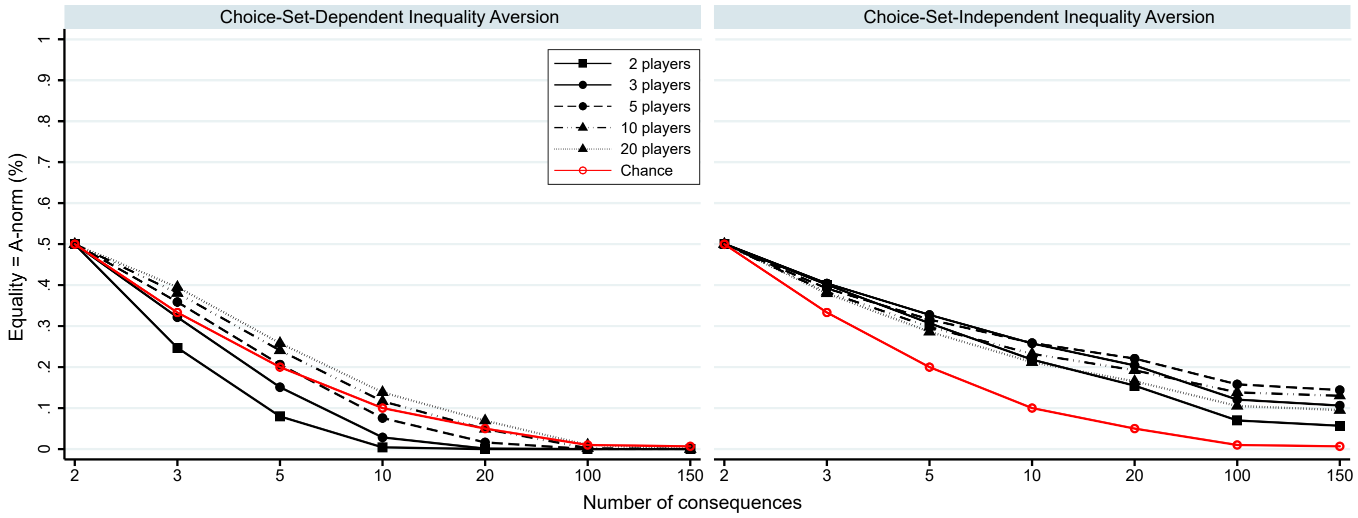


Figure 8: The percentage of D-norms that coincide with predictions of context-dependent and context-independent inequality aversion for random sets of payoff vectors. 500,000 sets for each case.

By way of contrast, we can also ask whether other well-known moral rules correspond to the D-norm in these settings. For example, if we repeat the exact same exercise as above and ask how often the D-norm makes the same prediction as inequality aversion in randomly chosen payoff vectors, we get rather different results. Figure 8 shows that for general random sets of payoff vectors, the context-dependent inequality aversion corresponds to the D-norm at a rate no better than chance, while the context-independent rule does only slightly better than chance. This is not surprising given that the D-norm is always Pareto optimal, whereas inequality aversion rules tend to favor payoff vectors close to the center of any set C. However, it is plausible that inequality aversion rules may be better suited to situations like the Dictator game in which payoff efficiency is constant. We explore the properties of the D-norm in those settings next.

5.2.2 Constant-Sum Settings

In the above example, we computed various injunctive norm functions 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 injunctive norms. So, the 80% result above should be taken with care.[13] 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 D-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, and the payoff efficiency moral rule does not have any explanatory power at all since all payoff vectors have the same payoff efficiency. We apply our method to choice settings of this kind, but this time ask whether the D-norm generally corresponds to moral rules that represent inequality aversion.

Figure 9 shows the percentage of cases (out of 500,000 randomly generated sets of payoff vectors from a simplex defined by the condition $\sum_{i\in N} u_i(x) = 1$ and by the condition that all payoffs are non-negative) in which the context-dependent and context-independent inequality aversion favor the same allocation as the D-norm. A first interesting observation is that both moral rules correspond to injunctive norm at a rate better than chance. However, the context-dependent rule does so more frequently than the context-independent one in settings with only a few consequences, though the gap shrinks as the choice set grows. Thus, we might expect the more complex context-dependent moral rule to be favored when the choice set is small (it fits better with injunctive norms) and the less complex context-independent

[13] 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.

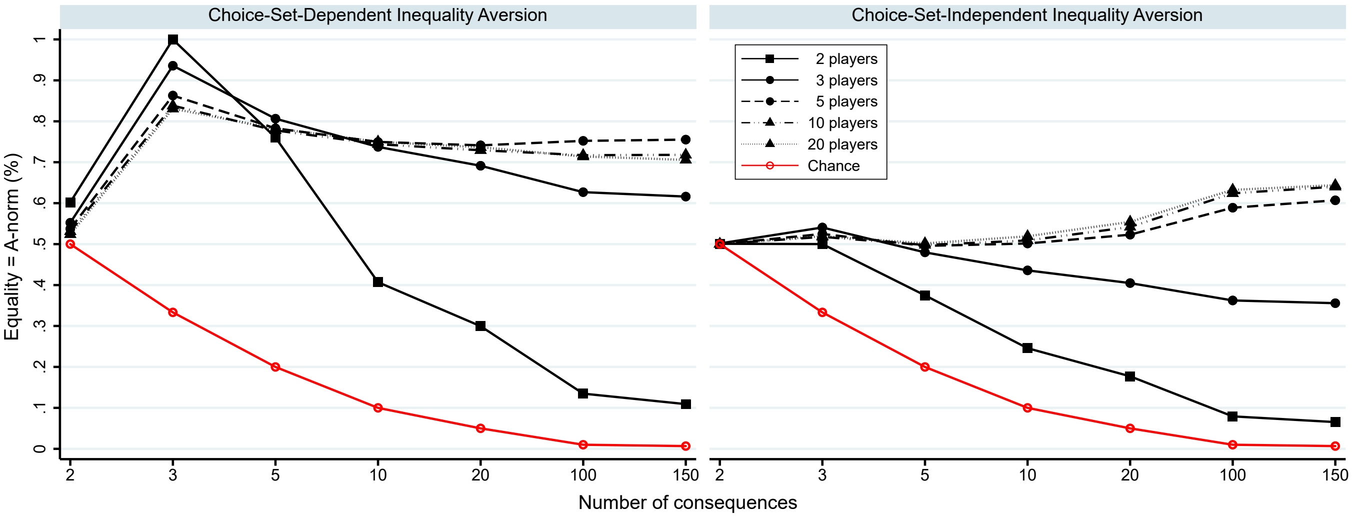


Figure 9: The percentage of D-norms that coincide with predictions of context-dependent and context-independent inequality aversion for random sets of payoff vectors with constant payoff efficiency. 500,000 sets for each case.

moral rule to be favored as the choice set grows large since it is about as likely to correspond to injunctive norms and is simpler to apply.

Conjecture 3. *In choice sets with constant or little varying payoff efficiency, inequality becomes a key force driving normative comparisons. Moral rules based on payoff equality are thus likely to be favored, with simpler (i.e. context-independent) versions of such rules more likely as the choice set grows.*

6 Meta-Theory of Moral Rules

The simulations above show that simple moral rules like payoff efficiency or inequality aversion can be good approximations of injunctive norms in certain classes of contexts. Given that moral rules are easier to articulate and use than injunctive norms, this gives them an "economic" advantage and suggests that moral rules may be wide-spread at least in the environments where it is inefficient to use injunctive norms. We thus show how KV's model of radically context-dependent injunctive norms can be used to provide a *meta-theory of moral rules*, predicting which moral rules are likely to emerge in which choice settings and helping to account why there exists a diverse set of moral rules used in different contexts.

In general, it is not hard to imagine that given some class of contexts (e.g., all Dictator games) and a population of players with specific characteristics (e.g., given some distributions of ϕ_i and cognitive abilities) we can determine 1) what is the expected norm-dependent utility from using injunctive norms in this class of games played by this population and 2) what is the expected norm-dependent utility from using some moral rule in the same conditions. These expected utilities can also include some estimates of costs of using both mechanisms. Such costs can be influenced for example by the (average) number of computations needed to compute the D-norm and the rule (see Section 4); the effects from the potentially easier spread of the rule as compared to the intuitions from injunctive norms; the effects of cognitive ability on costs of using the rule and D-norm, etc. Then, the expected utilities can be compared and a prediction made as to whether the moral rule will be adopted (if the expected utility of its usage is greater than that of injunctive norms) or not (the opposite). Roughly, moral rules will be preferred when the moral cost of their adoption (they only imprecisely reflect injunctive norms, so sometimes the rules can lead to morally undesirable consequences) is less than the cost of computation of injunctive norms.

If some process of this sort indeed takes place in reality—or that moral rules get adopted by populations when their usage brings higher norm-dependent utility than injunctive norms—then we can talk about the *evolution* of moral rules. Indeed, imagine a society living through a period of some change. For example, new technologies open new possibilities for profitable social interactions and thus new classes of contexts appear in which moral decision should be made. In such case, we can expect moral rules to evolve in this new environment that are the most efficient from the perspective of expected norm-dependent utility (given the characteristics of the population). Similarly, when some conditions change in existing classes of contexts and interactions (new laws get passed, new products appear, new cultural traits get introduced) moral rules can be expected to change as well, evolving to better reflect injunctive norms and to make cooperation more efficient.

7 Discussion

Although the meta-theory of moral rules proposed above just outlines a general mechanism of adoption of moral rules, it can have many interesting implications. For example, the idea that computing moral intuitions is costly suggests straight away that we should expect to see moral rules used extensively in situations where people are involved in many social interactions of the same type (e.g., queueing, making moral judgements in typical criminal cases, buying goods). When there are many such typical interactions, moral rules become attractive because even a small decrease in the cost of each transaction can have a significant impact over time.

To give an example, consider the evolution of law. When people live in a small community where everyone knows each other and there is little bad behavior, there is no need to have a complex system of rules (laws) that determine what exact punishment should be given for specific crimes. A small community has a luxury to consider each case of misbehavior separately using injunctive norms that can provide the best possible moral solution. However, as the community grows the number of crimes goes up as well and we obtain a situation where courts need to deal with many typical cases where similar moral decisions should be made. Here, it is not inconceivable that rules for dealing with specific classes of cases (e.g., murder) will develop to streamline the process.

Another similar example can be found in retail. Why is it that when we come to a grocery store in a developed country we see the price of some item and we pay it without question? Why cannot we bargain about the price in the store as is common in many other parts of the world? It may be that some people are poorer than others and can only pay a smaller price. However, such moral concerns do not enter our minds in this specific situation. One reason might be that such “unquestioned price-taking” has evolved into a moral rule in the specific circumstances and populations of the developed countries. First, grocery stores serve many customers each day, so having fast transactions is crucial in the sense that grocery stores simply do not have time to discuss the price of each item with each customer. Second, the populations in the developed world are known to trust strangers and behave morally overall, which implies that they might believe that the price is reasonable and thus might be more likely to accept it; in addition, there are consumer protection laws, agencies, etc. that monitor the situation with prices; as an outcome, grocery store customers can be sure that the price they are supposed to “take” is a reasonable one, which makes it easier for them to agree to this specific moral rule. So, the characteristics of the population and the existing institutions can also influence the type of moral rules that emerge.

Overall, the meta-theory of moral rules that we propose in this paper can be a valuable tool to study the emergence, change, and disappearance of moral rules that can have a large and tangible effect on economic activity. The understanding that moral judgements are costly allows us to have a different perspective on the patterns of moral rules that we observe in reality. It becomes possible to uncover deeper reasons for why certain moral rules exist and which policy to use to change them.

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.

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

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

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

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.

Kimbrough, E. and Vostroknutov, A. (2023). 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.

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

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

Wilson, B. J. (2020). What is right is not taken out of the rule, but let the rule arise out of what is right. In *The Property Species: Mine, Yours, and the Human Mind*. Oxford University Press.

Appendix (for online publication)

A Proofs

Proof of Proposition 1. $(1 \Rightarrow 2)$. By R3, $f(0,0) = 0$ 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 R3 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. $\triangle$

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

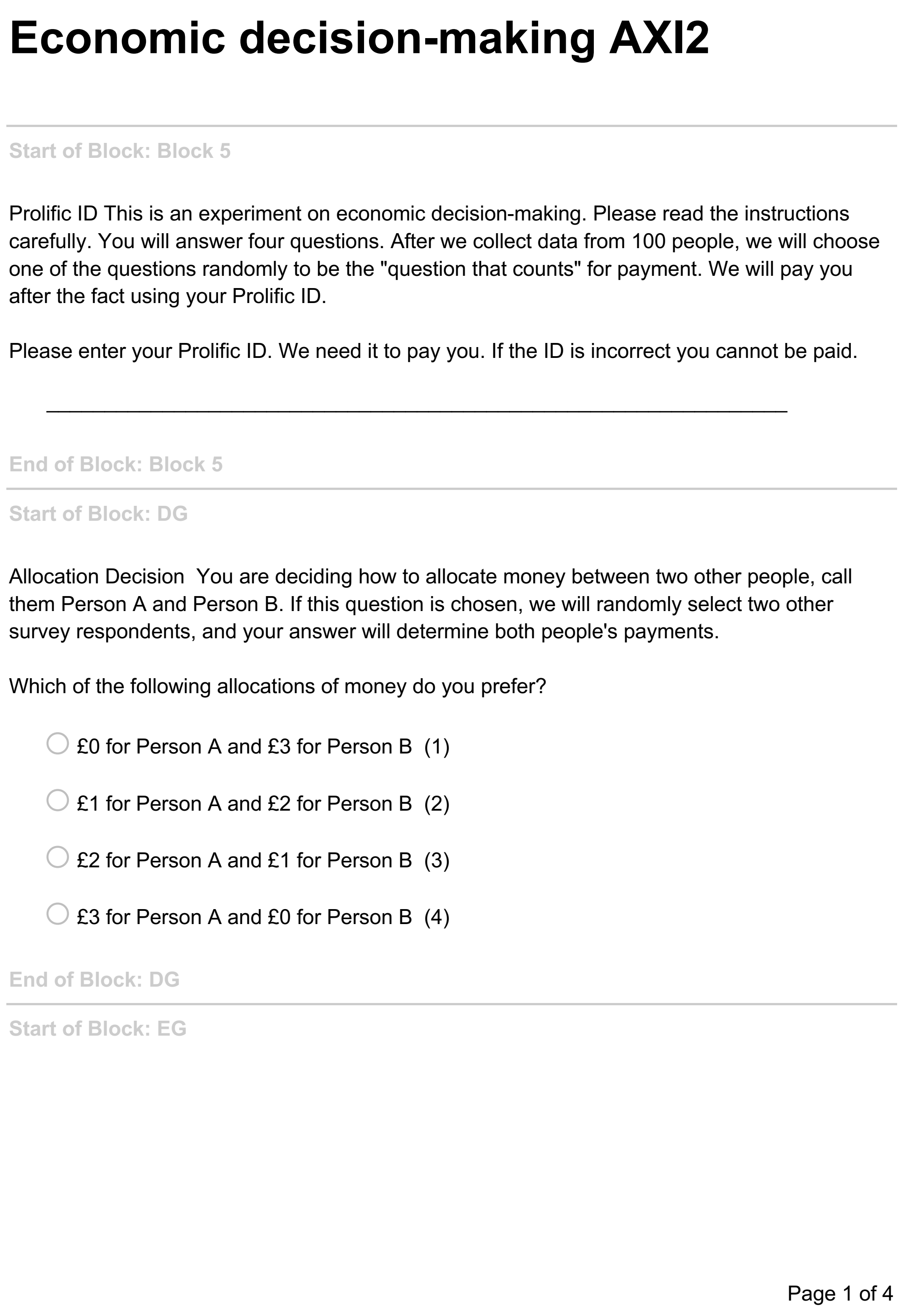
Proof of Proposition 3. $(1 \Rightarrow 2)$. For all $F_1, \ldots, F_N \in \mathbb{R}_+$, R7 implies $E(F_1, \ldots, F_N) = E(0, \ldots, 0) + \sum_{i \in N} \omega_i F_i$. By R6, $G(F_1, \ldots, F_N) = \sum_{i \in N} \omega_i F_i$. Thus, since F_i satisfy R5 and f_i satisfy R1-R4, we have

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

as desired. $\triangle$

$(2 \Rightarrow 1)$. R5-R7 are trivial. We get R1-R4 from the proof of Propositions 1. $\square$

B Experiment Details



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

Page 1 of 4

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