What matters and how it matters: A choice-theoretic representation of moral theories

Dietrich, Franz and List, Christian

Paris School of Economics, CNRS, London School of Economics

February 2016

What Matters and How it Matters: A Choice-Theoretic Representation of Moral Theories*

Franz Dietrich
CNRS & Paris School of Economics

Christian List
London School of Economics

23 February 2016

Abstract

We present a new "reason-based" approach to the formal representation of moral theories, drawing on recent decision-theoretic work. We show that any moral theory within a very large class can be represented in terms of two parameters: (i) a specification of which properties of the objects of moral choice matter in any given context, and (ii) a specification of how these properties matter. Reason-based representations provide a very general taxonomy of moral theories, as differences among theories can be attributed to differences in their two key parameters. We can thus formalize several important distinctions, such as between consequentialist and non-consequentialist theories, between universalist and relativist theories, between agent-neutral and agent-relative theories, between monistic and pluralistic theories, between atomistic and holistic theories, and between theories with a teleological structure and those without.

1 Introduction

The aim of this paper is to propose a new approach to the formal representation of moral theories. We show that any moral theory within a very large class can be represented in terms of two parameters:

*Earlier versions of this work were presented at a Workshop on "Social Dynamics of Information Change", University of Amsterdam, December 2013, at the Second International Conference on Economic Philosophy, Strasbourg, October 2014, at the Seminar "Sciences, Normes, Décisions", Université Paris-Sorbonne, May 2015, at the Workshop on "Ethics, Uncertainty, and Decision Theory", Australian National University, July 2015, and at the Conference on "Deontic Logic and Ethics", Venice Interrnational University, November 2015. We thank the participants at these events for their comments. We are very grateful to John Broome, Seth Lazar, Thomas Scanlon, Kai Spiekermann, and Laura Valentini for helpful conversations and/or written feedback.

(i) a specification of which properties of the objects of moral choice matter in any given context, and

(ii) a specification of how these properties matter.

The first parameter tells us what the normatively relevant properties are, the second how these properties (or sets of them) are to be "weighed" relative to one another. We call a representation of a moral theory in terms of these parameters a *reason-based representation*; we give a precise definition below. Reason-based representations encode not only a theory's *action-guiding recommendations* (i.e., how we should act, according to the theory), but also the *reasons* behind those recommendations (i.e., why we should act in that way).

As will become clear, reason-based representations provide a very general taxonomy of moral theories, since differences among theories can be attributed to differences in their two key parameters. In this way, we can capture a number of important distinctions, such as between consequentialist and non-consequentialist theories, between universalist and relativist theories, between agent-neutral and agent-relative theories, between monistic and pluralistic theories, between atomistic and holistic theories, and between theories with a teleological structure and those without.

Reason-based representations also shed light on an important, but still under-appreciated phenomenon: different moral theories may coincide in all their action-guiding recommendations, despite arriving at them in different ways. Put differently, the same action-guiding recommendations may be explained in more than one way.[1] For example, some deontologists and some consequentialists may agree on all "ought" statements, but offer different explanations for them. Our reason-based approach allows us to investigate this phenomenon, which we call the *underdetermination of moral theory by deontic content*.

In developing our approach, we build on the existing debate on whether all moral theories can be "consequentialized".[2] Roughly speaking, a moral theory is "consequentializable" if its action-guiding recommendations are the same as those of some "counterpart theory" that is structurally consequentialist. Some scholars, such as Jamie Dreier, suggest that every moral theory – or at least every plausible one – can be represented in a consequentialist format, provided we employ a sufficiently broad notion of *consequences*. Others, such as Campbell Brown, argue that "[t]here are in fact limits to consequentialization": whenever a moral theory does not satisfy certain formal constraints, it defies

[1]For discussions of this point, see, e.g., Dreier (1993), Broome (2004, ch. 3), and Portmore (2011).

[2]See ibid. and Brown (2011), Dreier (2011), and Tenenbaum (2014).

consequentialization, in a sense that can be made precise.[3]

We agree that there are limits to consequentialization, unless we permit unilluminating ways of redescribing the options of moral choice. This raises the question of whether a theory that falls outside those limits can be represented in some other canonical way. Our framework gives a positive answer to this question. And even when a theory *can* be consequentialized, the framework allows us to represent not only the action-guiding recommendations but also the underlying reasons. Furthermore, we can go beyond the debate on consequentialization and ask, for each attribute A in the set {consequentialist, universalist, agent-neutral, monistic, atomistic, teleological}, which theories have that attribute and which ones are at least "A-izable", i.e., redescribable in a form that has attribute A. We conclude the paper by showing how our framework can deal with moral choices involving uncertainty. Our analysis builds on earlier work on reason-based choice, but the paper is self-contained.[4]

2 What do we mean by a moral theory?

We begin with some basic terminology and informal background to our discussion. We define our central concepts more precisely in the subsequent formal exposition.

2.1 Normative versus axiological theories

Moral theories, broadly construed, can be of at least two kinds: they can be *axiological* or *normative*. An *axiological theory* is a theory of how good or bad (or better or worse relative to one another) certain objects of assessment are. The objects of assessment can be, for instance, possible worlds, states of affairs, actions, or consequences. A *normative theory*, by contrast, is a theory of which actions (or policies, plans, arrangements) are permissible or impermissible, right or wrong. Action-guidance is usually delivered by normative theories, not by axiological ones.

Of course, there can be connections between theories of these two kinds. Many normative theories are based, directly or indirectly, on an axiological theory. Consequentialist theories, such as utilitarianism, are like this. They are normative theories that are defined on the basis of an underlying axiological theory which refers to the goodness of consequences (their sum-total utility in the case of utilitarianism). Typically, they deem actions permissible or right if and only if they bring about the best

[3]See, respectively, Dreier (1993, 2011) and Brown (2011, p. 750).

[4]We build on the formal framework in Dietrich and List (2016), offering a normative rather than positive-explanatory interpretation. An earlier, more distant precursor is Dietrich and List (2013).

feasible consequences. As already mentioned, the "consequentialization debate" concerns the question of whether all normative theories can be re-expressed in this format, under a suitable interpretation of "best consequences".

More broadly, a theory is *teleological* if its criterion for the permissibility of actions is that they are the best feasible ones, according to *some* underlying axiological theory, though not necessarily one that focuses on consequences alone. Assessments of goodness or betterness could focus, for instance, on how the acts relate to the context of choice. We can then say that a normative theory can be *teleologized* if it can be re-expressed in a teleological format, i.e., if we can construct a teleological counterpart theory with the same action-guiding recommendations.[5]

We here focus on normative theories, rather than axiological theories. Therefore, when we speak of a moral theory, this should be understood to refer to a normative theory, unless otherwise stated.

2.2 Normative theories and their deontic content

Any normative theory – perhaps together with some auxiliary assumptions – entails a body of permissibility verdicts: verdicts about which actions are permissible in any given context, and which are not. These are the theory's action-guiding recommendations. Let us call this the *deontic content* of the theory.[6]

Typically, the theory itself is more than just an enumeration of its permissibility verdicts. It goes beyond its deontic content. This is because the theory offers a systematization or explanation of the implied permissibility verdicts, for example by identifying the reasons and general principles underpinning them. It is entirely possible, for instance, that different normative theories entail the same permissibility verdicts, despite arriving at them in different ways. Several authors have recognized this phenomenon, sometimes under the label of "extensional equivalence".[7] A striking suggestion of extensional equivalence can be found in Derek Parfit's recent book, *On What Matters*.[8] Parfit argues that his favourite versions of consequentialism, Kantianism, and Scanlonian contractualism essentially coincide in their recommendations and can be seen as attempts to climb the same mountain from different sides. Similarly, John Broome observes that the same normative recommendations may be derived from different axiologies.[9]

[5]For an earlier discussion of these issues, see Broome (2004, ch. 3).
[6]This follows Brown's (2011) and Portmore's (2011) terminology (*deontic outputs* or *verdicts*).
[7]See, e.g., Portmore (2009) and Dreier (2011).
[8]See Parfit (2011).
[9]See, e.g., Broome (2004, ch. 3).

If we are interested not only in how we ought to act, but also in why we ought to act in that way, then we cannot generally consider two extensionally equivalent theories as equivalent *simpliciter*. As Douglas Portmore observes:

> "[E]ven if two theories agree as to which acts are right, that does not mean that they agree on what makes those acts right... [M]oral theories are in the business of specifying what makes acts right. And so even two moral theories that are extensionally equivalent in their deontic verdicts can constitute distinct moral theories – that is, distinct theories about what makes acts right."[10]

The present problem is analogous to the case of science, where two or more distinct theories may explain the same observations and thus be observationally equivalent, despite being explanatorily different: W. V. Quine's famous *empirical underdetermination problem*.[11] Proponents of an instrumentalist view of science typically deny that there is much at stake in our choice among observationally equivalent theories. They think that the main point of a scientific theory is to accommodate the empirical observations; our theoretical constructs are just instrumentally useful representation devices.[12] By contrast, scientific realists insist that there is a fact of the matter as to which theory offers the right explanation of the observations: no more than one of the rival theories can be true. Here, it is not only the theory's observable implications than can be true or false, but also the theoretical constructs offered as an explanation: the unobservables.

Similarly, normative theories are underdetermined by their deontic content. Some scholars, such as Jamie Dreier, think that there is not much at stake in such cases of underdetermination.[13] This view parallels the instrumentalist one in science. However, we think that those who attach significance to normative reasons ought to disagree. From their perspective, an accurate representation of a moral theory should capture not only the theory's deontic content, but also the underlying reasons or principles.

3 How do we formalize a theory's deontic content?

We first explain how to formalize a theory's deontic content, following John Broome and Campbell Brown in taking a decision-theoretic approach, and then we briefly revisit the limits of consequentialization.[14]

[10]See Portmore (2011, p. 109). Dreier (2011) also discusses this passage.
[11]See Quine (1975) and, for a more recent discussion, List (1999).
[12]See, e.g., van Fraasen (1980).
[13]See, e.g., Dreier (2011).
[14]See, in particular, Broome (1991) and Brown (2011).

3.1 Choice contexts, options, and rightness functions

As noted, a theory's action-guiding recommendations are encoded by its deontic content: a specification of which actions are permissible in each context, and which are not. We formalize this as follows.

Let $\mathcal{K}$ be a set of possible *choice contexts* that an agent may be faced with. Each context K (an element of $\mathcal{K}$) is a situation in which the agent has to choose among, or appraise, some options, such as different actions or prospects. Let $[K]$ denote the set of options in context K. It is most natural to interpret these as the "feasible" options in context K, where readers may plug in their preferred notion of feasibility. On a thinner interpretation, we may interpret the options in $[K]$ simply as those that are being appraised in that context.[15] For ease of exposition, we call the options in $[K]$ the "available" ones in context K. The set $[K]$, in turn, is a subset of a universal set X of possible options.[16]

For each context K, a normative theory specifies which of the available options are permissible, and which not. To capture this, we introduce the notion of a *rightness function*. This is a function, denoted R, which assigns to each context K the set $R(K)$ of "permissible" or "right" options in that context, where $R(K)$ is a subset of $[K]$. In decision-theoretic terms, the function R is a *choice function*, reinterpreted to capture "permissible" or "right" choice, rather than "actual" or "formally rational" choice.[17]

A rightness function expresses a theory's deontic content. If, according to the theory, every available option is permissible in context K, then the set of permissible options $R(K)$ coincides with $[K]$. If there is a unique permissible option in K, then $R(K)$ is singleton. If there is no permissible option, then $R(K)$ is empty, in which case the agent faces a moral dilemma. We call the rightness function R *dilemma-free* if $R(K)$ is non-emtpy for every context K. (It is worth noting one complication: some theories distinguish not only between permissible and impermissible options, but also between "merely permissible" options and "supererogatory" ones, which are beyond the call of duty. We set this complication aside, though our framework can be extended to

[15] The first interpretation of $[K]$, as the set of "feasible" options, is natural if the normative theory in question obeys an "ought implies can" constraint. The second interpretation is compatible with the idea that we may appraise options as permissible or impermissible even before assessing their feasibility.

[16] In particular, the set X contains every option that is available in at least one possible context. Formally, $X \supseteq \cup_{K \in \mathcal{K}}[K]$.

[17] Formally, R is a function from the set $\mathcal{K}$ of contexts into a set of subsets of X, where, for each K, $R(K) \subseteq [K]$. Note that a standard choice function assigns to every context a *non-empty* set of chosen options. We do not require non-emptiness.

accommodate it.[18])

The present formalism permits a variety of interpretations. Contexts can be specified as richly as needed for an adequate description of the choice situation, and even the agent's identity can in principle be built into the notion of a context; more on this in Section 5.3.[19] Similarly, options can be specified in a variety of ways, though it is best not to pack contextual features into the specification of the options themselves. Contextual features should be included in the specification of the contexts in which those options are available. It will then be possible, at least theoretically, to encounter the same option in more than one context. This makes it meaningful to ask questions such as the following: "would option x, which is permissible in context K, still be permissible in a different context K'?" Later, we explicitly distinguish between properties that options have intrinsically and properties they have in relation to the context.

3.2 Consequentialization revisited

Our goal is to find a canonical way of systematizing the permissibility verdicts encoded by a given rightness function. One approach to this problem, familiar from the literature on consequentialization, is to try to identify a binary relation over the options, typically interpreted as a *betterness relation*, such that, for any context, the permissible options are the highest-ranked available options according to that relation. Formally, given a rightness function R, we are looking for a binary relation $\succeq$ on the set X such that, for any context K,

$$R(K) = \{x \in [K] : x \succeq y \text{ for all } y \in [K]\},$$

[18]Some options among the permissible ones may stand out as "saintly" or "heroic" (using the terms of Urmson 1958). We can capture this idea by introducing a *saintly or heroic choice function* H which assigns to each context K the set $H(K)$ of those options that a saint or hero would choose, subject to the constraint that $H(K) \subseteq R(K)$. For some contexts K, $H(K)$ may be a *proper* subset of $R(K)$ (i.e., $H(K) \subsetneq R(K)$), so that we can interpret the options in $H(K)$ as the supererogatory ones. Here the saint or hero does something that goes beyond the call of duty. For other contexts K, $H(K)$ and $R(K)$ may coincide. A saint's or hero's choices will not always differ from those of an "ordinary" moral agent; in some contexts, there is simply no scope for supererogation. An option is *supererogatory* in context K if it is in $H(K)$ *and* $H(K) \subsetneq R(K)$. There is some debate about whether the phenomenon of supererogation exists. In the present terms, this debate concerns the question of whether there is an H function as distinct from the R function. For discussion, see Heyd (2015).

[19]To make this explicit, we could define a context K as a pair $\langle Y, \Phi \rangle$ of (i) a set Y of available options (i.e., $[K] = Y$) and (ii) a set Φ of other contextual features (such as the time, the past history, certain background facts, the cultural environment, or, in principle, even the agent's identity).

where "$x \succeq y$" means "x is ranked at least as highly as y according to the relation $\succeq$".[20] When there exists such a relation, we say that the rightness function is *representable by a binary relation*. Furthermore, when the relation is transitive, it is conventional to call the representation *structurally consequentialist*. In fact, some philosophers, such as John Broome, hold that a binary relation deserves the name "betterness relation" *only if* it is transitive. A normative theory is *consequentializable* if its rightness function admits a structurally consequentialist representation.

When can a rightness function be represented in this way? Since rightness functions are formally the same as choice functions, we can bring a large body of work in decision theory to bear on this question (in line with Campbell Brown's analysis).[21] In particular, there are well-known necessary and sufficient conditions for the representability of a choice function by a binary relation, under a variety of constraints on that relation. These results apply equally to rightness functions. In Appendix A, we state one illustrative such representation theorem, which gives us a precise dividing line between those rightness functions that can be represented by a binary relation, and those that cannot. The key point is that *if*, *and only if*, a rightness function satisfies a particular structural condition, it can be represented by a binary relation.

3.3 Two problems

There are at least two problems with the attempt to systematize or explain rightness functions by representing them in terms of binary relations over the options. The first is that some reasonable rightness functions cannot be represented in this way. For a simple example (due to Amartya Sen), consider a rightness function that encodes norms of politeness.[22] When you are offered a choice between different pieces of cake at a dinner party, politeness commands that you do not choose the biggest piece, because that would be greedy. So, when a big (x), a medium-sized (y), and a small piece of cake (z) are available, it is permissible to choose any of the three, except the biggest. In particular, choosing the medium-sized piece is perfectly ok. Formally, if $[K] = \{x, y, z\}$, then $R(K) = \{y, z\}$. However, when the big piece is unavailable (so you are now choosing between the two smaller pieces), then choosing the medium-sized piece (y) is no longer permissible, because it is now the biggest on offer. Formally, if $[K'] = \{y, z\}$, then $R(K') = \{z\}$. No binary relation over the different pieces of cake could represent this

[20] The relation $\succeq$ also induces a strict relation $\succ$ and an indifference relation $\sim$. For any x and y, we have $x \succ y$ if and only if $x \succeq y$ and not $y \succeq x$, and we have $x \sim y$ if and only if $x \succeq y$ and $y \succeq x$.

[21] See Brown (2011).

[22] See Sen (1993).

rightness function. For a binary relation to represent it, it would have to rank z strictly above y and also not do so.[23]

Of course, one might try to respond to this problem by redescribing the options in a richer way, but this response is problematic too. If we build the entire choice context into the description of the options, so that each option can occur in only one context, we can trivially represent *any* rightness function in terms of an artificially constructed binary relation (as shown formally in Appendix A), but that binary relation will be completely uninformative. It will simply be a cumbersome redescription of the rightness function itself, enumerating all its recommendations in a relation-theoretic format. Moreover, it will still be true that the rightness function *as originally defined* admits no representation in terms of a binary relation.

The second problem with the attempt to explain rightness functions in terms of binary relations is a version of the uninformativeness problem just mentioned. Such a representation – even when it exists – tells us very little about the *reasons* underpinning the permissibility verdicts encoded by that rightness function. Suppose we are asked: "why is x permissible and y is not?" If we simply say "because x is better than y", this is not very informative. It would be legitimate to ask a further question, namely: "what is it about x and y that makes x better than y?" A representation of a rightness function in terms of a binary relation is silent on that further question. We would like go beyond merely enumerating "brute" goodness facts or betterness facts; we would like to say something about each option's right-making or wrong-making features.

4 Reason-based representations

As announced, our idea is to represent any moral theory in terms of two parameters:

(i) a specification of which properties of the options matter in any given context, and

(ii) a specification of how these properties matter.

To make this precise, we first introduce the notion of a *property* and give a taxonomy of different kinds of properties. We then define the notion of a *reasons structure*, which is our formalization of (i) and (ii). We finally explain how a reasons structure entails, and thereby explains, a rightness function.[24]

[23] We have to explain (1) $R(K) = \{y, z\}$ and (2) $R(K') = \{z\}$. A necessary condition for explaining (1) is $y \sim z$, while a necessary condition for explaining (2) is $z \succ y$. We cannot have both.

[24] The formalism draws on the positive, decision-theoretic framework in Dietrich and List (2016).

4.1 Properties

At a first gloss, a property is a feature that an option may or may not have, so that a property picks out the set of those options that have that property. For example, if the options are possible meal choices, the property *vegetarian* picks out the set of those meals that involve no meat. As will become clear, however, this definition is not sufficiently general.

Instead of defining properties as features of options *simpliciter*, we define them as features of option-context pairs. An *option-context pair* is a pair of the form $\langle x, K \rangle$, where x is an option (an element of X) and K is a context (an element of $\mathcal{K}$). Formally, a *property* is a primitive object P that picks out a set of option-context pairs, called the *extension* of P and denoted $[P]$. Whenever a pair $\langle x, K \rangle$ is contained in $[P]$, this means that option x has property P in context K. Sometimes we also say: the option-context pair $\langle x, K \rangle$ has property P. A property P may be of three different kinds:[25]

- P is an *option property* if its possession by an option-context pair depends only on the option, not on the context.[26]
- P is a *context property* if its possession by an option-context pair depends only on the context, not on the option.[27]
- P is a *relational property* if its possession by an option-context pair depends on both the option and the context.[28]

Suppose, for example, that X is a set of meal choices, and $\mathcal{K}$ is a set of menus from which one may choose. The property *vegetarian* is an option property. If a particular meal option is vegetarian in the context of one menu, then it is still vegetarian in the context of another. The property *offering two or more options* is a context property. Whether an option-context pair has that property depends solely on the context (here the menu), irrespective of the particular option we are considering. Finally, the property *most calorific* is a relational property, since the same meal option can be most calorific relative to one menu, but not relative to another. Likewise, properties such as *polite* or *norm-conforming* are relational properties, since the same act can be polite or norm-conforming in one context, but not in another. Think about the difference between public and private contexts or between contexts involving different cultures.

[25]In what follows, we will only consider properties whose extension is "non-trivial", in that it is neither empty nor total: i.e, at least one, but not all, option-context pairs have the property in question.

[26]Formally, for all x in X and all K, K' in $\mathcal{K}$, $\langle x, K \rangle \in [P]$ if and only if $\langle x, K' \rangle \in [P]$.

[27]Formally, for all K in $\mathcal{K}$ and all x, x' in X, $\langle x, K \rangle \in [P]$ if and only if $\langle x', K \rangle \in [P]$.

[28]Formally, it is neither an option property nor a context property.

We write $\mathcal{P}$ to denote the set of those properties that may be candidates for normatively relevant properties: we call these the *admissible properties*. If there are no constraints on the properties that might turn out to be normatively relevant according to some moral theory, then $\mathcal{P}$ could, in principle, be the universal set of all logically possible properties (which contains at least one property for every possible extension). If, on the other hand, some properties are so far-fetched that they could never be normatively relevant, then $\mathcal{P}$ could be more restricted.[29]

4.2 The notion of a reasons structure

We are now able to define the notion of a reasons structure. It specifies what the normatively relevant properties in each context are, and how these properties (or sets of them) are to be "weighed" relative to one another. Formally, a *reasons structure* is a pair $\mathcal{R} = \langle N, \geq \rangle$ consisting of:

- A *normative relevance function* N, which assigns to each context $K \in \mathcal{K}$ a set $N(K)$ of *normatively relevant* properties in that context.[30]
- A *weighing relation* $\geq$ over sets of properties, formally a binary relation whose relata are subsets of $\mathcal{P}$. When this relation ranks one set of properties above another, this can be interpreted to mean that the first property set is at least as "good", or at least as "choice-worthy", as the second.[31]

For example, in the case of a utilitarian theory, the function N assigns to every context the set of all "utility properties"; and the relation $\geq$ ranks sets of such properties in terms of the criterion that more utility is better than less; we make this more precise below.

To regiment our formalism, we impose one invariance constraint on the normative relevance function N: whenever two contexts K and K' have the same context properties, then the same properties are normatively relevant in those contexts.[32] By contrast, we

[29] The set $\mathcal{P}$ can be partitioned into a subset of option properties, a subset of context properties, and a subset of relational properties, denoted $\mathcal{P}_{\text{option}}$, $\mathcal{P}_{\text{context}}$, and $\mathcal{P}_{\text{relational}}$. Some of these subsets might be empty. For any option x and any context K, we further write $\mathcal{P}(x, K)$ for the set of all properties of $\langle x, K \rangle$ (among the properties in $\mathcal{P}$); $\mathcal{P}(x) = \mathcal{P}(x, K) \cap \mathcal{P}_{\text{option}}$ for the set of all option properties of x; and $\mathcal{P}(K) = \mathcal{P}(x, K) \cap \mathcal{P}_{\text{context}}$ for the set of all context properties of K.

[30] Formally, N is a function from $\mathcal{K}$ into $2^{\mathcal{P}}$.

[31] As is standard, we write $>$ and $\equiv$ for the strict and indifference relations induced by $\geq$.

[32] The idea behind this constraint is the following: if we assert that different properties are normatively relevant in contexts K and K', then we should be able to point to some features of those contexts – some context properties – to which this difference can be attributed. If two contexts are property-wise

initially impose no restrictions on the weighing relation $\geq$. In principle, it could fall well short of the requirements of an ordering; it might even be intransitive. However, if one wishes to interpret the weighing relation as a "betterness relation" and accepts the view that "betterness" is transitive, then one might restrict attention to transitive weighing relations. Our next step is to explain how a reasons structure entails permissibility verdicts.

4.3 The entailed rightness function

In any given context K, the question is: which options among the available ones are permissible, according to a given reasons structure $\mathcal{R}$? To answer this question, we look at the available options through the lens of their normatively relevant properties, according to the normative relevance function N. For each option x and each context K, we write $\mathcal{P}(x, K)$ to denote the set of *all* properties of this option-context pair (among the properties in $\mathcal{P}$). Since $N(K)$ is the set of all *normatively relevant* properties in context K, the normatively relevant properties of option x in context K are obviously those that lie in the intersection of $\mathcal{P}(x, K)$ and $N(K)$, i.e., those properties of $\langle x, K\rangle$ that are also in $N(K)$, formally

$$\mathcal{P}(x, K) \cap N(K).$$

Let us write $N(x, K)$ to denote this set. We then assess different available options by comparing their sets of normatively relevant properties, using the weighing relation $\geq$. More precisely, option x is deemed at least as choice-worthy as option y in context K if and only if the weighing relation ranks the set $N(x, K)$ at least weakly above the set $N(y, K)$. Now the permissible or "right" options are the ones that are at least as choice-worthy as all other available options, formally:

$$R(K) = \{x \in [K] : N(x, K) \geq N(y, K) \text{ for all } y \in [K]\}.$$

We call the function R thus defined the *rightness function entailed by the reasons structure* $\mathcal{R}$.[33]

It is helpful to give some informal illustrations. Returning to our example of a utilitarian theory, here the set $N(K)$ of normatively relevant properties is always the set of all "utility properties", i.e., all properties of the form "the total utility of the option's consequences is such-and-such". The weighing relation $\geq$ then ranks property sets in

indistinguishable, then it is hard to explain how they could give rise to different normatively relevant properties. Formally, if $\mathcal{P}(K) = \mathcal{P}(K')$, then $N(K) = N(K')$.

[33]To refer to this rightness function, we sometimes also use the notation $R^{\mathcal{R}}$.

terms of total utility; for instance, the singleton property set {"the total utility is 15"} is ranked above the set {"the total utility is 10"}. In each context K, the function R now selects the utility-maximizing option(s) among the available ones.

Next consider a simple deontological theory which defines permissibility in terms of the minimization of rights violations. Here, in each context, the set $N(K)$ consists of all "rights-violation properties" in terms of which options are to be assessed. For any property P in $N(K)$, an option's having that property in context K means that, by choosing that option in that context, the agent would violate some right. The weighing relation may rank sets of properties by size: i.e., one set of properties is ranked above a second set if the first set is smaller than the second (i.e., it consists of fewer rights-violation properties than the second). For a stricter version of this theory, only the empty set of properties – standing for no rights violations – is ranked above every set of properties, while no other sets of properties are weakly or strictly ranked above others. According to this version of the theory, only options that involve no rights violations are ever deemed choice-worthy (so $R(K)$ can sometimes be the empty set).

We say that a rightness function has a *reason-based representation* if there exists some reasons structure that entails it.[34] In Appendix B, we show that every rightness function within a very large class can be represented in this way, including even rightness functions that defy consequentialization in the conventional sense. In fact, the more properties we are willing to invoke in constructing a reason-based representation, the more rightness functions we are able to explain. If the set $\mathcal{P}$ of admissible properties is the universal set of all logically possible properties, then every logically possible rightness function – however far-fetched – can be *formally* represented by some reasons structure. Of course, far-fetched rightness functions may be representable only in a rather gerry-mandered way, and they may no longer be representable if we impose some reasonable constraints on the properties we are willing to invoke.

5 A taxonomy of moral theories

We propose to represent a moral theory canonically as a rightness function that is underwritten by some reasons structure. We now show that this way of representing moral theories yields a very general taxonomy. Specifically, we characterize moral theories in terms of six key distinctions. The first four refer to a theory's normative relevance function; the last two refer to its weighing relation.

[34] Formally, R has a reason-based representation if there exists some reasons structure $\mathcal{R} = \langle N, \geq \rangle$ such that $R = R^{\mathcal{R}}$.

5.1 Consequentialist versus non-consequentialist theories

The distinction between structurally consequentialist and non-consequentialist theories can be drawn in terms of the nature of the properties that these theories deem normatively relevant. On a simple definition, a theory is *consequentialist* if, according to that theory, only consequences matter;[35] it is non-consequentialist if, according to it, some things other than consequences matter, at least sometimes. We formalize this as follows.

A theory with reasons structure $\mathcal{R} = \langle N, \geq \rangle$ is *structurally consequentialist* if it never deems any properties other than option properties normatively relevant; i.e., for every context K, $N(K)$ consists of option properties alone. The theory is *structurally non-consequentialist* if it sometimes deems relational or context properties normatively relevant; i.e., for some context K, $N(K)$ includes some relational or context property. We also call such properties *context-related properties*.

One might quarrel whether our distinction between option properties and context-related properties matches the standard distinction between "consequence properties" and "non-consequentialist properties". However, we suggest that any doubts on this front should lead us, not so much to challenge our distinction, but rather to specify the options in X and the contexts in $\mathcal{K}$ more carefully. A plausible requirement is that, for any option-context pair $\langle x, K \rangle$, x should encode everything that is deemed to belong to the consequences of choosing x, while K should encode only the situation in which the choice takes place, including the range of alternative options. If our specification of the options and contexts meets this requirement, then our distinction between option properties and context-related properties will be aligned with the distinction between "consequence properties" and "non-consequentialist properties".

Some examples help to illustrate our distinction between consequentialist and non-consequentialist theories. Suppose the options are welfare distributions across a society of n individuals. So, each option is some n-tuple of the form $\langle w_1, w_2, ..., w_n \rangle$, where w_i is the welfare level of the i^{th} individual. The agent might be a social planner choosing among these options. Consider three theories.

- **Utilitarianism:** This says that we should choose a distribution $\langle w_1, w_2, ..., w_n \rangle$ which maximizes the *total welfare*, defined as $w_1 + w_2 + ... + w_n$.
- **Entitlement satisfaction:** Here, in each context K, there is an n-tuple of entitlements, $\langle e_1, e_2, ..., e_n \rangle$, where e_i is the welfare level to which the i^{th} individual is entitled (perhaps on grounds of effort or desert); the n-tuple of entitlements

[35] Sometimes a distinction is drawn between *causal* and *constitutive* consequences of an action; see, e.g., Dreier (2011). Our formal analysis below is ecumenical.

may differ from context to context. Now the theory says that we should choose a distribution $\langle w_1, w_2, ..., w_n\rangle$ that maximizes the number of individuals whose entitlements are satisfied, i.e., for whom $w_i \geq e_i$.[36]

- **Satisficing utilitarianism:** This says that we should choose a distribution $\langle w_1, w_2, ..., w_n\rangle$ in which the total welfare is at least 0.8 times as great as the total welfare in every available alternative.[37]

All of these theories can be easily represented in our framework. To represent utilitarianism, we must invoke properties of the form

$$P_{\text{wel}=w}: \text{ “The total welfare is } w\text{”},$$

where w is some real number. Call such properties *total-welfare properties*. The reasons structure is $\mathcal{R} = \langle N, \geq\rangle$, where:

- for every context K, $N(K)$ is the set of all total-welfare properties;
- the weighing relation is defined over singleton sets consisting of total-welfare properties, where $\{P_{\text{wel}=w}\} \geq \{P_{\text{wel}=w'}\}$ if and only if $w \geq w'$.

To represent the entitlement-satisfaction theory, we must invoke properties of the form

$$P_i: \text{ “The entitlement of individual } i \text{ is satisfied”},$$

where i ranges over the n individuals in the society under consideration. Call such properties *entitlement-satisfaction properties*. The reasons structure is $\mathcal{R} = \langle N, \geq\rangle$, where:

- for every context K, $N(K)$ is the set of all entitlement-satisfaction properties;
- the weighing relation is defined as follows: for any two sets of entitlement-satisfaction properties – call them S and S' – we have $S \geq S'$ if and only if S contains at least as many entitlement-satisfaction properties as S'.[38]

[36] If there are ties between distributions under this criterion, we might add some tie-breaking criterion, for instance one that requires minimizing the sum-total shortfall.

[37] Brown (2011) and Dreier (2011) discuss this as an example of a non-consequentialist theory.

[38] If we wanted to include the tie-breaking criterion mentioned in a previous footnote, we would have to add to each set $N(K)$ all properties of the form $P_{\text{short}=\delta}$ (“the total shortfall is δ”), where δ is some real number, and the *total shortfall* for any distribution $\langle w_1, w_2, ..., w_n\rangle$ relative to $\langle e_1, e_2, ..., e_n\rangle$ is defined as $\sum_{i=1,...,n} \min(0, e_i - w_i)$. The weighing relation would have to be defined as follows: $S \geq S'$ whenever (i) S contains more entitlement-satisfaction properties than S', *or* (ii) S and S' contain the same number of entitlement-satisfaction properties and $\delta \leq \delta'$, where $P_{\text{short}=\delta} \in S$ and $P_{\text{short}=\delta'} \in S'$.

Finally, to represent satisficing utilitarianism, we must invoke the following property:

P_{satisf} : "The total welfare is at least 0.8 times as great as in every alternative".

Call this the *satisficing property.* The reasons structure is simply $\mathcal{R} = \langle N, \geq \rangle$, where:

- for every context K, $N(K)$ consists of the satisficing property alone;
- the weighing relation ranks $\{P_{\text{satisf}}\}$ above the empty set.

It is easy to see that utilitarianism is consequentialist in the sense we have defined, while the other two theories are not. This is because the normatively relevant properties according to utilitarianism, total-welfare properties, are option properties, while the normatively relevant properties according to the other theories (entitlement-satisfaction properties and the satisficing property) are not. Whether a welfare distribution has a particular total-welfare property – it offers such-and-such total welfare – depends only on the welfare distribution itself. By contrast, whether the individuals' entitlements are satisfied depends on the n-tuple of entitlements in the context: it is a relational property. Likewise, whether a distribution's total welfare is at least 0.8 times as great as that in every alternative depends on the other available distributions. Along similar lines, theories that deem some "essentially comparative" properties normatively relevant, as famously suggested by Larry Temkin, qualify as non-consequentialist.[39] We give an example later.

Importantly, we have not built transitivity of the weighing relation into our definition of consequentialism. Although a transitive ordering of consequences is sometimes considered one of the defining characteristics of consequentialism, the question of whether the normatively relevant properties are restricted to option properties is orthogonal to the question of whether the weighing relation is transitive. We take the former question to pick out the difference between consequentialism and non-consequentialism, and the latter to pick out the difference between teleology and non-teleology; more on this later.

5.2 Universalist versus relativist theories

Informally, universalism can be understood as the view that what matters is always the same, regardless of the context in which the assessment takes place, while relativism can be understood as the view that different things matter in different contexts. For example, different cultural contexts or social practices may make different properties normatively relevant. Later we comment on a special form of relativism, *agent-relativity.*

[39] See, e.g., Temkin (1996, 2012).

Our framework offers a natural formalization of the distinction between universalism and relativism. A theory with reasons structure $\mathcal{R} = \langle N, \geq \rangle$ is *structurally universalist* if the set of normatively relevant properties $N(K)$ is the same across all contexts; i.e., the normative relevance function N is constant. The theory is *structurally relativist* if $N(K)$ is not the same across all contexts; i.e., N is non-constant. Different contexts render different properties normatively relevant.

The three theories discussed in the last subsection, utilitarianism, entitlement satisfaction, and satisficing utilitarianism, are all examples of structurally universalist theories; they each specify a constant set of normatively relevant properties: the set of all total-welfare properties, the set of all entitlement-satisfaction properties, and the satisficing property, respectively. If we constructed a theory that deems welfare properties relevant in some contexts and entitlement-satisfaction properties in others, this would be structurally relativist. Communitarians sometimes endorse structurally relativist theories, insofar as they take the normatively relevant properties to depend on social conventions and social meanings in the contexts in question.

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">Context-variant?</td></tr>
<tr><td>No</td><td>Yes</td></tr>
<tr><td rowspan="2">Context-related?</td><td>No</td><td>Universalist consequentialism (e.g., utilitarianism)</td><td>Standard relativism (e.g., practice-dependent theories)</td></tr>
<tr><td>Yes</td><td>Univ. non-consequentialism (e.g., entitlement satisfaction, deontological theories)</td><td>Relativist deontology (e.g., cultural-norms-based theories)</td></tr>
</table>

Table 1: Two kinds of context-dependence

It is worth noting that the two distinctions we have drawn so far, namely between consequentialism and non-consequentialism and between universalism and relativism, each concern the question of whether a theory's reasons structure is context-dependent in a particular way. The first distinction (between consequentialism and non-consequentialism) concerns the question of whether the normatively relevant properties include context-related properties. Call a reasons structure *context-related* if the answer to this question is positive, and *context-unrelated* otherwise. The second distinction (between universalism and relativism) concerns the question of whether or not the normatively relevant properties vary across contexts. Call a reasons structure *context-variant* if the answer to this question is positive, and *context-invariant* otherwise. Table 1 shows how these two distinctions can be combined.

5.3 Agent-neutral versus agential-relative theories

Roughly speaking, a moral theory is agent-neutral if the identity of the agent makes no difference to its prescriptions (other things being equal), while it is agent-relative if it does. Utilitarianism is the standard example of an agent-neutral theory, while ethical egoism – which recommends that each agent should pursue his or her own self-interest – is a familiar example of an agent-relative one: its action-guiding recommendations depend on who the agent is. There are many ways of making the distinction between agent-neutrality and agent-relativity precise.[40]

To explicate the distinction in our framework, we first revisit the notion of a context. When we introduced that notion, we said that contexts can be specified as richly as needed for an adequate description of the choice situation, and we acknowledged that even the agent's identity can be built into the context. We may thus think of a context K as a triple $\langle i, Y, \Gamma \rangle$, where i is the agent, Y the set of available options (i.e., $[K] = Y$), and Γ a set of other situational or environmental features.[41] We can now refine our taxonomy of two kinds of context-dependence from the end of the last subsection.

We begin by reconsidering the distinction between context-invariant and context-variant reasons structures, our dividing line between universalist and relativist theories. Recall that a reasons structure is context-invariant if the normatively relevant properties – those in $N(K)$ – do not vary across contexts, and context-variant if they do. If we take each context K to be a triple $\langle i, Y, \Gamma \rangle$, we can refine this distinction by asking

(i) whether $N(K)$ varies with changes in the agent i;[42]

(ii) whether $N(K)$ varies with changes in the set of available options, Y;[43] and

(iii) whether $N(K)$ varies with changes in the other situational or environmental features, as specified by Γ.[44]

The answer to question (i) yields the distinction between *agent-variant* and *agent-invariant* reasons structures; the answer to question (ii) yields the distinction between *menu-variant* and *menu-invariant* reasons structures (where the "menu" is the set of available options); and the answer to question (iii) yields the distinction between *situation/environment-variant* and *situation/environment-invariant* reasons structures.

[40] See, e.g., Ridge (2011).

[41] In an earlier footnote, we suggested thinking of a context as a pair $\langle Y, \Phi \rangle$ of a set of available options and a set Φ of other contextual features. We can interpret Φ as subsuming i and Γ.

[42] Formally, are there $K=\langle i, Y, \Gamma \rangle$ and $K'=\langle i', Y', \Gamma' \rangle$, with $Y=Y'$ and $\Gamma=\Gamma'$, such that $N(K) \neq N(K')$?

[43] Formally, are there $K=\langle i, Y, \Gamma \rangle$ and $K'=\langle i', Y', \Gamma' \rangle$, with $i=i'$ and $\Gamma=\Gamma'$, such that $N(K) \neq N(K')$?

[44] Formally, are there $K=\langle i, Y, \Gamma \rangle$ and $K'=\langle i', Y', \Gamma' \rangle$, with $i=i'$ and $Y=Y'$, such that $N(K) \neq N(K')$?

A theory that deems different properties normatively relevant depending on the social or cultural context exhibits situation/environment-variance. A theory that instructs different agents to focus on different properties exhibits agent-variance. Ethical egoism, for instance, may be an illustration, deeming only individual i's welfare properties normatively relevant for individual i. Accordingly, we might define *agent-neutrality* as the absence of agent-variance in a theory's reasons structure, and *agent-relativity* as the presence of such agent-variance. Since agent-variance is a special case of context-variance, agent-relativity is thus a special form of relativism.

However, there is also another way of distinguishing between agent-neutral and agent-relative theories. To see this, let us go back to the distinction between context-unrelated and context-related reasons structures, our dividing line between consequentialist and non-consequentialist theories. Recall that a reasons structure is context-unrelated if the normatively relevant properties – those in $N(K)$ – are always option properties, while it is context-related if they include context-related properties, at least sometimes. And recall that a property P is context-related if its possession by an option-context pair depends, at least in part, on the context.[45] If we interpret each context K as a triple $\langle i, Y, \Gamma \rangle$, we can refine these definitions too:

(i) a property P is *agent-related* if its possession by an option-context pair depends, at least in part, on the agent i;[46]

(ii) a property P is *menu-related* if its possession by an option-context pair depends, at least in part, on the set of available options, Y;[47]

(iii) a property P is *situation/environment-related* if its possession by an option-context pair depends, at least in part, on the other situational or environmental features, as specified by Γ.[48]

Call a reasons structure *agent-related*, *menu-related*, and *situation/environment-related* if the normatively relevant properties in $N(K)$ include, respectively, agent-related, menu-related, and situation/environment-related properties, at least for some K. The reasons structure of utilitarianism is free from any such properties. By contrast, the reasons

[45] Formally, for some x in X and some K, K' in $\mathcal{K}$, $\langle x, K \rangle \in [P]$ and $\langle x, K' \rangle \notin [P]$.

[46] Formally, for some x in X and some $K = \langle i, Y, \Gamma \rangle, K' = \langle i', Y', \Gamma' \rangle$ in $\mathcal{K}$ with $Y = Y'$ and $\Gamma = \Gamma'$, $\langle x, K \rangle \in [P]$ and $\langle x, K' \rangle \notin [P]$.

[47] Formally, for some x in X and some $K = \langle i, Y, \Gamma \rangle, K' = \langle i', Y', \Gamma' \rangle$ in $\mathcal{K}$ with $i = i'$ and $\Gamma = \Gamma'$, $\langle x, K \rangle \in [P]$ and $\langle x, K' \rangle \notin [P]$.

[48] Formally, for some x in X and some $K = \langle i, Y, \Gamma \rangle, K' = \langle i', Y', \Gamma' \rangle$ in $\mathcal{K}$ with $i = i'$ and $Y = Y'$, $\langle x, K \rangle \in [P]$ and $\langle x, K' \rangle \notin [P]$.

structure of the entitlement-satisfaction theory is situation/environment-related, as the entitlement-satisfaction properties are situation/environment-related. (Whether someone's entitlements are satisfied depends on what he or she is entitled to in the given situation.) The reasons structure of satisficing utilitarianism is menu-related, since the satisficing property is menu-related: whether the sum-total welfare in a given distribution is at least 0.8 times as great as that in every alternative depends on the available "menu" of distributions. However, none of these reasons structures deem any agent-related properties normatively relevant. We might define *agent-neutrality* as the absence of agent-relatedness in a theory's reasons structure, and *agent-relativity* as the presence of agent-relatedness. Defined in this way, agent-relativity is a special form of non-consequentialism, not of relativism.

To illustrate these two orthogonal ways of defining the distinction between agent-neutrality and agent-relativity, consider ethical egoism, stated in the setup of our earlier examples where the options are welfare n-tuples of the form $\langle w_1, w_2, ..., w_n\rangle$. We now assume that the agent is one of the n individuals. Ethical egoism requires individual i to choose a distribution which maximizes w_i. According to one reason-based representation, the theory is agent-relative in the relativist sense (rather than the non-consequentialist one). Here, we invoke properties of the form

$$P_{\text{wel}_i=w}: \text{ "Individual } i\text{'s welfare is } w\text{"},$$

where i is some individual and w is some real number. Call these the *welfare properties for individual i*. Now the reasons structure is $\mathcal{R} = \langle N, \geq\rangle$, where:

- for every context K, $N(K)$ is the set of all welfare properties for the individual i named in $K = \langle i, Y, \Gamma\rangle$.
- the weighing relation is defined over singleton sets consisting of individual welfare properties, where $\{P_{\text{wel}_i=w}\} \geq \{P_{\text{wel}_i=w'}\}$ if and only if $w \geq w'$.

Since $N(K)$ varies with the agent, the reasons structure is agent-variant and thus context-variant, our defining condition of relativism. At the same time, $N(K)$ contains only option properties. Thus the reasons structure is context-unrelated, our defining condition of consequentialism.

Contrast this with a second representation of ethical egoism, according to which the theory is agent-relative in the non-consequentialist sense. Here, we invoke properties of the form

$$P_{\text{agent's wel}=w}: \text{ "The agent's welfare (or perhaps 'my' welfare) is } w\text{"},$$

where w is some real number. Call these the *agent-centred welfare properties.* Now the reasons structure is $\mathcal{R} = \langle N, \geq \rangle$, where:

- for every context K, $N(K)$ is the set of all agent-centred welfare properties;
- the weighing relation is defined over singleton sets consisting of agent-centred welfare properties, where $\{P_{\text{agent's wel}=w}\} \geq \{P_{\text{agent's wel}=w'}\}$ if and only if $w \geq w'$.

Note that $N(K)$ does not vary with the identity of the agent. Regardless of who makes the choice, the theory always deems all agent-centred welfare properties normatively relevant. In fact, the reasons structure is completely context-invariant, and thus universalist as we have defined it. At the same time, the properties contained in $N(K)$, agent-centred welfare properties, are agent-related: options do not have these properties "instrinsically", but only in relation to the agent in question.[49] The reasons structure is therefore context-related, our defining condition of non-consequentialism.

In sum, we have identified two ways of drawing the line between agent-neutral and agent-relative theories. We can view agent-relativity either as a special form of relativism, or as a special form of non-consequentialism.

5.4 Monistic versus pluralistic theories

Like the previous distinctions, the next one concerns a theory's normative relevance function. Moral philosophers distinguish between monistic and pluralistic theories. Roughly, monistic theories assess options in terms of a single normatively relevant property, while pluralistic theories assess them in terms of multiple normatively relevant properties, which may need to be traded off against one another. In our framework, a theory with reasons structure $\mathcal{R} = \langle N, \geq \rangle$ is *structurally monistic* if each option has precisely one normatively relevant property in each context; i.e., $N(x, K)$ is singleton for every option-context pair $\langle x, K \rangle$. The theory is *structurally pluralistic* if some options have two or more normatively relevant properties in some contexts; i.e., there exists at least one option-context pair $\langle x, K \rangle$ for which $N(x, K)$ contains more than one property.[50]

Some of our illustrative theories – e.g., utilitarianism, satisficing utilitarianism, and ethical egoism – are monistic in this sense. By contrast, the entitlement-satisfaction theory and our deontological theory that defines permissibility in terms of minimal

[49] Formally, whether an option-context pair $\langle x, K \rangle$ has the property $P_{\text{agent's wel}=w}$ depends on both x and the agent i in the triple $\langle i, Y, \Gamma \rangle = K$.

[50] For the purpose of distinguishing between monism and pluralism, we set aside the limiting case in which $N(x, K)$ is empty, i.e., where x has no normatively relevant properties in context K.

rights violations are pluralistic; they take options to have multiple normatively relevant properties. Similarly, we obtain a pluralistic theory if we add side-constraints to utilitarianism, requiring the maximization of sum-total welfare, subject to some additional requirements. Options would then have more than one normatively relevant property, such as *their sum-total welfare being such-and-such* and *the side-constraint being satisfied/violated*. For another example, consider a functioning-or-capability theory, which assesses each option in terms of the bundle of functionings or capabilities that it generates. Here, each option could have multiple functioning-or-capability properties in a given context, and they could all be normatively relevant.

While the present distinction between structural monism and structural pluralism refers to each option's *number* of normatively relevant properties, one could revise (or strengthen) the definition of monism by requiring that the normatively relevant properties of all options (and in all contexts) be *of the same kind*. So $N(K)$ would only ever include properties of a single kind: e.g., only welfare properties, or only rights properties, or only duty-properties, and so on. To make this precise, we would have to introduce an equivalence relation over properties, capturing sameness of kind, and demand that, for a monistic theory, the normatively relevant properties always fall into the same equivalence class.

Finally, we can also represent those pluralistic theories that associate pluralism with moral dilemmas, by deeming some options mutually incomparable in light of their disparate properties. We do not require that the weighing relation be complete (where *completeness* means that no sets of properties are left unranked). It is possible, then, that some sets of properties are mutually incomparable. If those properties turn out to be the normative relevant properties of some available options, this may generate a situation in which the rightness function deems no options permissible.[51]

5.5 Teleological versus non-teleological theories

While the distinctions we have drawn so far refer to a theory's normative relevance function, we now turn to some distinctions that refer to the weighing relation. We begin with the distinction between teleological and non-teleological theories. As already noted, a theory is conventionally called *teleological* if its criterion for the permissibility of options is that they are the best feasible ones, according to *some* underlying axiology, though not necessarily an axiology that focuses on consequences alone.

In our framework, this requires that the weighing relation over sets of properties admit a "betterness" interpretation. If we accept the view that betterness is transitive,

[51] For further discussion of pluralism, see Mason (2015).

then such an interpretation is available only in those cases where the weighing relation is transitive. We thus call a theory with reasons structure $\mathcal{R} = \langle N, \geq \rangle$ *teleological* if $\geq$ is transitive, and *non-teleological* otherwise.

All the examples of theories that we have discussed so far are teleological in this sense. But even some theories that may at first sight seem non-teleological – because of their non-consequentialism – can be expressed in a teleological format. Larry Temkin's work offers some examples (though Temkin himself would interpret them differently, namely as evidence for the intransitivity of betterness). Let us here give a Temkin-inspired example that has been discussed by Alex Voorhoeve. In this example, the set X consists of three options (quoting and adapting Voorhoeve's wording):[52]

terminal: curing one young person of a terminal illness;

considerable: saving 10000 people from a considerable impairment;

slight: saving several billion people from a very slight impairment.

We assume, for the sake of the example, that the total-welfare properties of these options are as follows. The welfare of *terminal* is 5000; the welfare of *considerable* is 10000; and the welfare of *slight* is 15000. So, the three options have the option properties $P_{\text{wel}=5K}$, $P_{\text{wel}=10K}$, and $P_{\text{wel}=15K}$, respectively. Let us grant the following Temkin-inspired intuitions:

- When faced with a choice between *terminal* and *considerable*, we ought to choose *considerable*, on total-welfare grounds.

- When faced with a choice between *considerable* and *slight*, we ought to choose *slight*, again on total-welfare grounds.

- When faced with a choice between *terminal* and *slight*, we ought to choose *terminal*. Given the dramatic difference between "death" and "very slight impairment", it would be disrespectful not to save the one from death for the sake of saving the billions from a very slight impairment.

We can make sense of these intuitions in a reason-based way. In addition to the total-welfare properties, we introduce a "disrespect property". An option x in context K is "disrespectful" (for short, D) if and only if there is at least one other available option y such that someone's welfare loss from choosing x rather than y is very significantly greater than anyone's welfare loss from choosing y rather than x. Suppose, for the sake

[52]See Voorhoeve (2013, p. 413).

of our example, that option *slight* has this property if and only if *terminal* is also among the available options; no other option among our three options has this property. Now consider the reasons structure $\mathcal{R} = \langle N, \geq \rangle$ where:

- for every context K, $N(K)$ consists of all the total-welfare properties and the disrespect property;
- the weighing relation ranks sets of properties as follows:

$$\{P_{\text{wel}=15K}\} > \{P_{\text{wel}=10K}\} > \{P_{\text{wel}=5K}\}$$
$$> \{P_{\text{wel}=15K}, D\} > \{P_{\text{wel}=10K}, D\} > \{P_{\text{wel}=5K}, D\}.$$

In a context in which only *terminal* and *considerable* are available, their sets of normatively relevant properties are $\{P_{\text{wel}=5K}\}$ and $\{P_{\text{wel}=10K}\}$, so that the theory recommends choosing *considerable* over *terminal.* In a context in which only *considerable* and *slight* are available, their sets of normatively relevant properties are $\{P_{\text{wel}=10K}\}$ and $\{P_{\text{wel}=15K}\}$, so that the theory recommends choosing *slight* over *considerable.* However, when both *terminal* and *slight* are available, their sets of normatively relevant properties are $\{P_{\text{wel}=5K}\}$ and $\{P_{\text{wel}=15K}, D\}$, and so the theory recommends choosing *terminal* over *slight*, in line with the Temkin-style intuitions. Interestingly, the weighing relation of this reasons structure is transitive, and so the theory counts as teleological, although it is clearly non-consequentialist. What is going on is simply that D is a relational property (it is "essentially comparative" in Temkin's terminology), and the option *slight* has this property in some contexts but not in others.

5.6 Atomistic versus holistic theories

Our final distinction is that between "atomistic" and "holistic" theories. Informally, a theory is *atomistic* if its assessment of any set of properties is simply the "sum" (or some other simple combination) of its assessments of the individual properties in this set, while a theory is *holistic* if this is not the case; we will make this more precise below. Atomistic theories are "generalist" insofar as the "valence" they give to any property is the same, irrespective of which other properties are present as well. A property that is at least *pro tanto* right-making within one combination of properties remains *pro tanto* right-making when combined with other properties. By contrast, holistic theories are "particularist", insofar as the "valence" they give to some properties may depend on the other properties they are combined with. According to such a theory, a property may be *pro tanto* right-making in the presence of one particular set of properties, and *pro tanto* wrong-making in the presence of another.

In our framework, the distinction between atomistic and holistic theories can be drawn in terms of the nature of those theories' weighing relations. We call a theory with reasons structure $\mathcal{R} = \langle N, \geq \rangle$ *atomistic* if the weighing relation $\geq$ is *separable*. Separability of $\geq$ means that, for any comparable subsets S, S' of $\mathcal{P}$, we have

$$S \geq S' \text{ if and only if } S \cup T \geq S' \cup T,$$

where T is any subset of $\mathcal{P}$ which does not overlap with S or S' but where $S \cup T$ and $S' \cup T$ are comparable.[53] The theory is *holistic* if its weighing relation is not separable. An important special case of separability is the *additively separable* case. Here, there exists a numerical *weighing function* w, which assigns to each property P in $\mathcal{P}$ a real number $w(P)$, its positive or negative "weight", such that, for any two subsets S, S' of $\mathcal{P}$,

$$S \geq S' \text{ if and only if } \sum_{P \in S} w(P) \geq \sum_{P \in S'} w(P).$$

Monistic theories, as we have defined them, are automatically atomistic,[54] but there are also pluralistic theories that are atomistic. For an easy construction of such a theory, we can simply define a positive or negative numerical weight for each normatively relevant property and then assess any set of properties by summing up their weights.

The kinds of "particularist" theories defended by Jonathan Dancy and others are holistic in the present sense. Dancy says:

> "A feature can make one moral difference in one case, and a different difference in another. Features have ... variable relevance. Whether a feature is relevant or not in a new case, and if so what exact role it is playing there (the 'form' that its relevance takes there) will be sensitive to other features of the case. This claim emerges as the consequence of the core particularist doctrine, which we can call the holism of reasons. This is the doctrine that what is a reason in one case may be no reason at all in another, or even a reason on the other side... [I]n some contexts the fact that something is against the law is a reason not to do it, but in others it is a reason to do it (so as to protest, let us say, against the existence of a law governing an aspect of private life with which the law should not interfere)."[55]

---

[53] Two sets of properties, such as S and S', are *comparable* if $S \geq S'$ or $S' \geq S$.

[54] To be precise, a monistic theory is atomistic, assuming (without loss of generality) that $\geq$ is defined only over singleton property bundles.

[55] See Dancy (2013).

What Dancy describes here entails holism as we have defined it (a non-separable weighing relation), but it also seems to entail relativism (a context-variant set of normatively relevant properties). Logically, one can have holism without relativism and vice versa: non-separability of the weighing relation and context-variance of the normatively relevant properties are independent from one another.

6 The moral underdetermination problem

Our reason-based framework confirms the existence of the moral underdetermination problem: the same rightness function will often admit more than one reason-based representation. Hence a theory's deontic content does indeed underdetermine the reasons structure underwriting it. This raises the question of how the six dimensions along which we have categorized moral theories manifest themselves in a theory's deontic content. Can we tell from the deontic content whether or not the underlying theory is consequentialist, universalist, agent-neutral, monistic, atomistic, and/or teleological? Or are some of these attributes "deontically inert"?

We address this issue by considering each of the six attributes and asking whether every rightness function – assuming it has a reason-based representation at all – admits a representation that has that attribute. This, in turn, sheds light on the question of whether, for every moral theory, there exists an extensionally equivalent counterpart theory with that attribute. So, we are asking, in effect, whether every moral theory can be "consequentialized", "teleoligized", "universalized", and so on. We focus on moral theories that are representable in our framework (i.e., whose rightness functions admit a reason-based representation, relative to the given set $\mathcal{P}$ of properties).[56]

6.1 Does every rightness function admit a consequentialist representation?

Here, the answer is "no". We obtain this negative answer not only if we define consequentialism in the traditional way – as representability in terms of a betterness ordering over the options – but also if we define it as we have proposed, namely in terms of the normative relevance of option properties alone. The reason for this negative answer is that, even when a rightness function has some reason-based representation, it need not have a reason-based representation in which only option properties are normatively

[56] When we refer to a rightness function in each of the following subsection titles, we are therefore, strictly speaking, referring to a *reason-based representable* rightness function.

relevant. The conditions for reason-based representability in a context-unrelated format (as surveyed in Table 1 above) are more demanding than those for reason-based representability *simpliciter*.[57] For instance, the norms of politeness in Amartya Sen's dinner-party example can be represented by a reasons structure in which the relational property *polite* is normatively relevant, while they cannot be represented by a reasons structure in which only option properties are relevant.

We would only be able to "consequentialize" every normative theory by redescribing the options themselves, but as already noted, this would not generally be very informative and would involve a significant departure from the original description of the moral choice problem. We may thus conclude that the distinction between consequentialist and non-consequentialist theories is deontically significant.

6.2 Does every rightness function admit a universalist representation?

Here, the answer is "yes", provided that (as assumed) the rightness function has some reason-based representation at all. Surprisingly, the conditions for the existence of a reason-based representation *simpliciter* are logically equivalent to the conditions for the existence of a reason-based representation without context-variant normative relevance.

Fact: A rightness function R has a reason-based representation with a constant normative relevance function if and only if it has a reason-based representation *simpliciter*.[58]

In other words, every reasons structure is *deontically equivalent* to some reasons structure that is context-invariant (relative to the same admissible set $\mathcal{P}$ of properties). Thus, every relativist moral theory does indeed have a universalist counterpart theory with exactly the same deontic content. In that sense, the distinction between universalist and relativist theories is really a distinction at the level of the reasons structure, not at the level of the deontic content. Crucially, however, the universalist counterpart theory of a relativist theory may have to be non-consequentialist: the price of avoiding context-variance is the normative relevance of context-related properties.

6.3 Does every rightness function admit an agent-neutral representation?

Here, the answer depends on which of our two orthogonal ways of drawing the distinction between agent-neutrality and agent-relativity we adopt. If we define agent-neutrality as

[57] See Dietrich and List (2016, Appendix A.3).

[58] This is a subtly stronger version of a result proved in Dietrich and List (2016, Appendix A.2). The earlier result – stated in terms of choice functions – requires the function R to be dilemma-free.

agent-invariance – so that agent-relativity becomes a special case of relativism – then the answer to our question is a qualified "yes". As we have seen, every rightness function that has *some* reason-based representation also has one with a constant normative relevance function; and so, *a fortiori*, it has one that is agent-invariant: the set of normatively relevant properties will not vary with the agent i. However, just as the typical cost of re-expressing a relativist theory in a structurally universalist format is context-relatedness in the normatively relevant properties, so the cost of the present exercise will be agent-relatedness in those properties. Our second representation of ethical egoism in Section 5.3 illustrates this. The reasons structure in that example is agent-invariant, but it deems agent-related properties normatively relevant.

If we define agent-neutrality as agent-unrelatedness – so that agent-relativity becomes a special case of non-consequentialism – then the answer to our question is "no": not every rightness function admits an agent-neutral representation. Just as it is not true that every non-consequentialist theory can be consequentialized – such that only option properties are normatively relevant – so the elimination of the normative relevance of agent-related properties will not generally be possible if the given theory has an agent-related reasons structure.

6.4 Does every rightness function admit a monistic representation?

Here, the answer is a qualified "no". As should be evident, for instance from our Temkin-inspired example, we can easily find some rightness functions which only admit pluralistic reason-based representations, at least relative to a given set $\mathcal{P}$ of admissible properties. Suppose, on the other hand, we impose no parsimony restrictions on the set $\mathcal{P}$ of admissible properties and include in $\mathcal{P}$ even properties of the form "being option x in context K". These are *maximally specific*, in that they have only a single option-context pair in their extension. We can then trivially represent every rightness function in a monistic format, as shown in Appendix B. The only normatively relevant property of each option x in each context K will be "being option x in context K". However, such a representation is monistic in a rather trivial sense and completely unilluminating. Given a more disciplined specification of the set $\mathcal{P}$, the distinction between monistic and pluralistic theories is deontically significant.

6.5 Does every rightness function admit a teleological representation?

Here, the answer is also a qualified "no" – although this "no" is less obvious than our negative answers to some of the previous questions. We have seen that a very large class of rightness functions – even the ones in Temkin-inspired examples – can be represented in terms of a reasons structure with a transitive weighing relation. Yet, if we hold the set $\mathcal{P}$ of admissible properties fixed, then the conditions for reason-based representability with a transitive weighing relation are logically more demanding than the conditions for reason-based representability *simpliciter*. In Appendix B, we give a formal example to illustrate this point. We are unsure, however, whether there are any plausible instances of moral theories that defy teleoligization. It is therefore possible that all or most *prominent* moral theories are representable in a teleological format.

6.6 Does every rightness function admit an atomistic representation?

Again, the answer is a qualified "no", although the point is similar to what we have said about monistic representations above. If we take the set $\mathcal{P}$ of admissible properties to be given, we may well have no choice but to represent a given rightness function in terms of a reasons structure with a non-separable weighing relation. If, on the other hand, we are free to specify the set $\mathcal{P}$ as permissively as we like and include even maximally specific properties in it, we can always construct an atomistic representation, which will coincide with the trivial monistic representation we have mentioned. Once we set this unconstrained case aside, however, the distinction between atomistic and holistic theories is indeed deontically significant.

7 Moral choices under uncertainty

We have argued that a moral theory can be represented as a rightness function underwritten by a reasons structure, and we have developed a taxonomy of theories on that basis. We have not discussed moral choices under uncertainty, and so one might wonder whether our framework can handle this case. We conclude by explaining the resources our framework offers for the assessment of options involving uncertainty. We discuss two ways of capturing uncertainty. The first is to take the options of choice to be *lotteries* (also known as "gambles" or "risky prospects"), i.e., probability distributions over outcomes. The second is to take the options to be *Savage acts*, i.e., functions from states of the world to outcomes, where there is uncertainty about the state of the world.[59]

[59]The two strategies follow, respectively, von Neumann and Morgenstern (1944) and Savage (1954).

7.1 Options as lotteries

We have not made any particular assumptions about the nature of the options that are available in each context K. They are simply drawn from some underlying set X of possible options. From a formal perspective, X could be any non-empty set. Just as it is admissible to take the options in X to be actions whose downstream consequences are fully known, so it is equally admissible to take those options to be lotteries.

To make this precise, let O be some non-empty set of outcomes: these could be payoffs, allocations of goods or welfare, or states of affairs, for example. A *lottery* is a probability distribution over O, formally a function that assigns to every outcome a non-negative number (its probability), with a sum-total of 1 across the outcomes.[60] Now let X be the set of all lotteries over O. Then, in each choice context K, the set $[K]$ of available options is some set of lotteries. A rightness function R, as before, assigns to each context the set of those options that are "right" or "permissible" in that context, i.e., it selects those lotteries that may be permissibly chosen.

To illustrate how we can represent some familiar normative theories of choice under uncertainty, consider two examples, each applied to the simple case where each element of the outcome set O is a particular welfare level that is attained under the given outcome.

- **Expected-welfare maximization:** This says that we should choose a lottery which maximizes the *expected welfare* (defined as the lottery's probability-weighted average of the welfare levels in the outcome set).

- **Maximin:** This says that we should choose a lottery for which the minimum welfare level whose probability is non-zero is maximal.

Each of these theories admits a straightforward reason-based representation. To represent expected-welfare maximization, we must invoke properties of the form

$$P_{\text{exp}=w}: \text{ "The expected welfare is } w\text{"},$$

where w is some real number. Call such properties *expected-welfare properties.* Now the reasons structure is $\mathcal{R} = \langle N, \geq \rangle$, where:

- for every context K, $N(K)$ is the set of all expected-welfare properties;

- the weighing relation is defined over singleton sets consisting of expected-welfare properties, where $\{P_{\text{exp}=w}\} \geq \{P_{\text{exp}=w'}\}$ if and only if $w \geq w'$.

[60] We here consider the simplest case, in which each lottery assigns non-zero probability to only finitely many outcomes. Of course, our analysis can be extended to measurable sets O.

To represent maximin, we must invoke properties of the form

$$P_{\text{min}=w}: \text{ “The minimum welfare whose probability is non-zero is } w\text{”},$$

where w is some real number. Call such properties *minimum-welfare properties*. Now the reasons structure is $\mathcal{R} = \langle N, \geq \rangle$, where:

- for every context K, $N(K)$ is the set of all minimum-welfare properties;
- the weighing relation is defined over singleton sets consisting of minimum-welfare properties, where $\{P_{\text{min}=w}\} \geq \{P_{\text{min}=w'}\}$ if and only if $w \geq w'$.

Both of those theories would count as consequentialist in our taxonomy, since the properties that matter according to them – expected-welfare properties and minimum-welfare properties – are option properties. Whether a lottery has a property such as $P_{\text{exp}=w}$ or property $P_{\text{min}=w}$ does not depend on the context.

We can define many other properties of lotteries, such as their variance, the difference between the best and the worst outcomes they may yield with non-zero probability, or the difference between their expected value and the average expected value across all the available lotteries. While the first two of these examples are still option properties, the last is a relational property; it depends not only on the lottery under consideration but also on the other available lotteries. Formally, the extension of any property in the present case is always the set of those lottery-context pairs that have the property.

Indeed, once we understand that options can be lotteries, and that those options-as-lotteries can have a variety of properties in each context – just like any other options – it should be clear that normative theories that deal with lotteries are just as easily representable as are theories for which options do not take this form. Our taxonomy also continues to apply. Indeed, it can be extended. To give just one example, one might distinguish between theories according to which only “modal” properties matter – i.e., properties that depend only on what can or cannot happen with non-zero probability – and theories according to which “probabilistic” properties matter as well – i.e., properties which depend also on the numerical probabilities of possible outcomes. The maximin theory is purely modal; expected-welfare maximization is a probabilistic theory.

7.2 Options as Savage acts

While lotteries explicitly encode the probabilities of outcomes, our second approach to capturing uncertainty is different. A *Savage act* is simply a specification of what will happen in each one of several possible states of the world, given that one performs the act. No assignment of probabilities is built into a Savage act.

Formally, we assume that there is a non-empty set S of possible states of the world; these are assumed to be beyond a given agent's control. The different states could be different weather conditions (e.g., whether there is a drought or not), different economic circumstances (e.g., the level of growth), different health conditions (e.g., who does or does not get ill), or any other external circumstances on which the outcome of an agent's actions may depend. As before, we assume that there is a set O of possible outcomes that might result from taking action. A *Savage act* is defined as a function from the set S into the set O. It assigns to each state of the world the outcome that would result from performing the act in that particular state. The set X of possible options then consists of all Savage acts, and as before, $[K]$ will always be some subset of X.

For a simple example, suppose that there are two states of the world: "flood" and "no flood". The Savage act of "building a flood protection barrier" would yield the outcome "moderate cost" in the state "no flood" (say, a payoff of -100), and the outcome "moderate cost and lucky escape" in the state "flood" (which, let us say, also amounts to a payoff of -100). The Savage act of "building no flood protection barrier", by contrast, would yield the outcome "no cost" in the state "no flood" (say, a payoff of 0), and the outcome "disaster" in the state "flood" (say, a payoff of -1000). Our assessment of those Savage acts may focus on a variety of properties, such as their worst or best possible outcomes, or the amount of "regret" that we would experience if our choice led to a bad outcome. Here are two illustrative theories:

- **Maximax:** This says that we should choose a Savage act with the greatest best-case scenario. In our simple example, this theory would recommend building no flood protection barrier.

- **Regret minimization:** This says that we should choose a Savage act which minimizes maximal regret across the different possible states of the world, where the *regret* in each state of the world is the discrepancy between the actual outcome in that state and the best possible outcome one could have attained through some available act. In our example, the maximal regret for "building a flood protection barrier" would be 100. This is the regret that we would experience in case no flood occurs. By contrast, the maximal regret for "building no flood protection barrier" would be 1000. This is the regret that we would experience in case a flood occurs. Thus the theory would recommend building a flood protection barrier.

Both of these theories can be represented in our framework. To represent maximax, we simply need to invoke properties of the form

$$P_{\text{max}=x}: \text{ "The maximum payoff is } x\text{"},$$

where x is some real number. Call such properties *best-case properties*. The reasons structure of the maximax theory is $\mathcal{R} = \langle N, \geq \rangle$, where:

- for every context K, $N(K)$ is the set of all best-case properties;
- the weighing relation is defined over singleton sets consisting of best-case properties, where $\{P_{\text{max}=x}\} \geq \{P_{\text{max}=x'}\}$ if and only if $x \geq x'$.

To represent the regret-minimization theory, we need to invoke properties of the form

$$P_{\text{reg}(s)=r}: \text{ "The regret in state } s \text{ is } r\text{"},$$

where r is some non-negative real number and s is some state of the world in S. Call such properties *regret properties*. Each act-context pair has as many regret properties as there are states in S: one property specifying what the regret would be in each state, if the act were performed. The reasons structure of the regret-minimization theory is $\mathcal{R} = \langle N, \geq \rangle$, where:

- for every context K, $N(K)$ is the set of all regret properties;
- the weighing relation is defined over sets of regret properties, where, for any two such sets A and B, we have $A \geq B$ if and only if

$$\max(r : P_{\text{reg}(s)=r} \in A \text{ for some } s \in S) \leq \max(r : P_{\text{reg}(s)=r} \in B \text{ for some } s \in S).$$

Again, the two theories can be readily categorized in our framework. For example, maximax is a consequentialist theory in our sense, since the best-case properties are option properties. An act's best-case outcome does not depend on the other available Savage acts. By contrast, regret minimization is a non-consequentialist theory. The regret properties of each act depend crucially on the other available acts, and so they are relational. However, like maximax, the regret-minimization theory is still universalist: it deems the same properties normatively relevant in all contexts.

As should be clear, there is absolutely no barrier to representing also more complicated normative theories of choice under uncertainty in the present framework.[61] Our point here has simply been to deliver a "proof of concept". A detailed analysis of reason-based choice under uncertainty is left for future work.

[61] One recent example of such a theory is Lara Buchak's risk-weighted expected utility theory (2013).

References

Bhattacharyya, A., P. K. Pattanaik, and Y. Xu (2011) "Choice, Internal Consistency and Rationality", *Economics and Philosophy* 27(2): 123–149.

Bossert, W., and K. Suzumura (2010) *Consistency, Choice, and Rationality*, Cambridge, MA (Harvard University Press).

Broome, J. (1991) *Weighing Goods: Equality, Uncertainty and Time*, Oxford (Blackwell).

Broome, J. (2004) *Weighing Lives*, Oxford (Oxford University Press).

Brown, C. (2011) "Consequentialize This", *Ethics* 121(4): 749–771.

Buchak, L. (2013) *Risk and Rationality*, Oxford (Oxford University Press).

Dancy, J. (2013) "Moral Particularism", *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), E. N. Zalta (ed.), available at:
http://plato.stanford.edu/archives/fall2013/entries/moral-particularism/

Dietrich, F., and C. List (2013) "A reason-based theory of rational choice", *Nous* 47(1): 104–134.

Dietrich, F., and C. List (2016) "Reason-based choice and context-dependence", *Economics and Philosophy.*

Dreier, J. (1993) "Structures of Normative Theories", *The Monist* 76(1): 22–40.

Dreier, J. (2011) "In Defense of Consequentializing", in M. Timmons (ed.), *Oxford Studies in Normative Ethics* 1, Oxford (Oxford University Press): 97–118.

Heyd, D. (2016) "Supererogation", *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), E. N. Zalta (ed.), available at:
http://plato.stanford.edu/archives/spr2016/entries/supererogation/

List, C. (1999) "Craig's Theorem and the Empirical Underdetermination Thesis Reassessed", *Disputatio* 7: 28–39.

Mason, E. (2015) "Value Pluralism", *The Stanford Encyclopedia of Philosophy* (Summer 2015 Edition), E. N. Zalta (ed.), available at:
http://plato.stanford.edu/archives/sum2015/entries/value-pluralism/

Parfit, D. (2011) *On What Matters*, Oxford (Oxford University Press).

Portmore, D. (2011) *Commonsense Consequentialism: Wherein Morality Meets Rationality*, Oxford (Oxford University Press).

Quine, W. V. (1975) "On Empirically Equivalent Systems of the World", *Erkenntnis* 9(3): 313–328.

Richter, M. (1971) "Rational Choice", in J. S. Chipman et al. (eds.), *Preferences, Utility, and Demand*, New York (Harcourt Brace Jovanovich): 29–58.

Ridge, M. (2011) "Reasons for Action: Agent-Neutral vs. Agent-Relative", *The Stanford Encyclopedia of Philosophy* (Winter 2011 Edition), E. N. Zalta (ed.), available at: http://plato.stanford.edu/archives/win2011/entries/reasons-agent/

Savage, L. (1954) *The Foundations of Statistics*, New York (Wiley).

Sen, A. K. (1993) "Internal Consistency of Choice", *Econometrica* 61(3): 495–521.

Temkin, L. S. (1996) "A Continuum Argument for Intransitivity", *Philosophy and Public Affairs* 25(3): 175–210.

Temkin, L. S. (2012) *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*, Oxford (Oxford University Press).

Tenenbaum, S. (2014) "The Perils of Earnest Consequentializing", *Philosophy and Phenomenological Research* 88(1): 233–240.

Urmson, J. O. (1958) "Saints and Heroes", in A. I. Melden (ed.), *Essays in Moral Philosophy*, Seattle (University of Washington Press): 198–216.

van Fraassen, B. C. (1980) *The Scientific Image*, Oxford (Oxford University Press).

von Neumann, J., and O. Morgenstern (1944) *Theory of Games and Economic Behavior*, Princeton (Princeton University Press).

Voorhoeve, A. (2013) "Vaulting Intuition: Temkin's Critique of Transitivity", *Economics and Philosophy* 29(3): 409–423.

Appendix A: When can a rightness function be represented by a binary relation?

A.1 An illustrative representation theorem

We limit ourselves to the statement of one illustrative representation theorem, adapted from Marcel Richter's work:[62]

Theorem 1. A rightness function R is representable by a binary relation $\succeq$ on X if and only if R satisfies the following axiom.

Axiom 1. For every context K in $\mathcal{K}$ and any available option x in $[K]$, if, for every available option y in $[K]$, there exists some context K' in $\mathcal{K}$ in which x is deemed right while y is available (i.e., $x \in R(K')$ and $y \in [K']$), then x is deemed right in K (i.e., $x \in R(K)$).

For the purposes of this paper, we need not worry too much about the interpretation of Axiom 1. It is simply a formal condition that some rightness functions satisfy and others violate. Its point is only to provide a dividing line between those rightness functions that can be represented by a binary relation, and those that cannot. Axiom 1 by itself does not guarantee representability of a rightness function R by a *transitive* binary relation, i.e., a betterness relation in the narrower sense. For the latter, stronger requirements on R are needed; the details need not concern us here.[63]

A.2 Consequentializing by redescribing the options

The limits to consequentialization established by results such as Theorem 1 depend on holding the set X of options fixed. If we permit redescriptions of the options, we can trivially represent any rightness function in terms of a binary relation. This formalizes Jamie Dreier's hypothesis that every set of action-guiding recommendations has some consequentialist representation, under a sufficiently permissive notion of consequences.[64]

Let R be *any* rightness function defined on $\mathcal{K}$. Formally, R maps each context K in $\mathcal{K}$ to a subset of $[K]$. We redescribe the options as follows. Construct a new set X'

[62] Richter (1971) proved the theorem for choice functions, the equivalents of dilemma-free rightness functions. We have verified that the result still holds if the "dilemma-free" restriction is dropped.

[63] For details, see Bossert and Suzumura (2010).

[64] Dreier (2011) calls this the "Extensional Equivalence Thesis", though he restricts it to the deontic content of "plausible" moral theories. Our analysis illustrates the generality of this hypothesis, from a purely formal perspective.

of options, defined as the set of all pairs of the form $\langle x, K\rangle$, where x is an option in X and K is a context in which x is available. For each context K, let us reinterpret the extension of K as the set

$$\{\langle x, K\rangle : x \in [K]\}.$$

We can now redefine the rightness function R as a function R' from $\mathcal{K}$ into the set of subsets of X'. For each K in $\mathcal{K}$, let

$$R'(K) = \{\langle y, K\rangle : y \in R(K)\}.$$

Then R' is trivially representable by a binary relation on X'. This is because each redescribed option in X' is available in only one context K; it is described so richly that its occurrence is not repeatable. To represent R', we simply need to construct a binary relation $\succeq$ on X' which ranks the permissible options in each context K strictly above the impermissible ones. Interpretationally, R' and the original rightness function R encode the same action-guiding recommendations.[65]

Appendix B: When does a rightness function have a reason-based representation?

B. 1 A representation theorem

As we will show, any rightness function that satisfies a relatively undemanding regularity axiom has a reason-based representation. To state this axiom, we require one preliminary definition. For any two sets of properties, S and S', we say that S *is sometimes deemed choice-worthy in the presence of* S' if there exists some context K in which some option instantiating S is deemed right while some option instantiating S' is available. (An option x is said to *instantiate* a set S of properties in a context K if $\mathcal{P}(x, K) = S$.)

Axiom 2. For every context K in $\mathcal{K}$ and any available option x in $[K]$, if, for every available option y in $[K]$, $\mathcal{P}(x, K)$ is sometimes deemed choice-worthy in the presence of $\mathcal{P}(y, K)$, then x is deemed right in K (i.e., $x \in R(K)$).

As in the case of Axiom 1 above, we should not worry too much about the interpretation of Axiom 2. It is simply a verifiable condition that a rightness function may or may not satisfy. The following result holds:

[65] The present construction is borrowed from Dietrich and List (2016). For an earlier, more sophisticated construction that establishes a similar point, see Bhattacharyya, Pattanaik, and Xu (2011).

Theorem 2. A rightness function R has a reason-based representation $\mathcal{R} = \langle N, \geq \rangle$ if and only if R satisfies Axiom 2.

Axiom 2 thus allows us to draw a line between those rightness functions that can be represented in a reason-based way and those that cannot. The present theorem is a reason-based counterpart of Theorem 1 above, which was essentially Richter's classic representation theorem.[66] Just as Theorem 1 only established conditions for the representability of a rightness function by *some* binary relation – not necessarily one with further properties such as transitivity – so Theorem 2 only establishes conditions for the representability of a rightness function by *some* reasons structure – not necessarily one with further properties such as transitivity of the weighing relation. The permissiveness of Theorem 2 is deliberate: the goal is precisely to arrive at a canonical way of representing a large class of possible rightness functions.

B.2 How permissive is this representation?

The permissiveness of our representation depends on how large the set $\mathcal{P}$ of admissible properties is: recall that $\mathcal{P}$ contains those properties that we admit as candidates for normatively relevant properties. The more properties we have at our disposal in constructing a reasons structure, the more rightness functions we can explain. Technically, the demandingness of Axiom 2 depends on how large or small the set $\mathcal{P}$ is.

In the limit, if $\mathcal{P}$ contains all logically possible properties, then every logically possible rightness function has a reason-based representation. To see this, suppose, in particular, that the set $\mathcal{P}$ contains every property of the form:

$$P_{x,K}: \text{ ``The option is } x \text{, and the context is } K\text{''},$$

where x is an option and K a context.[67] Such a property is called *maximally specific*, as it is only satisfied by a single option in a single context. We can then represent any logically possible rightness function in our framework, namely by constructing the following reasons structure $\mathcal{R} = \langle N, \geq \rangle$:

- for every context K, $N(K)$ is the set of *all* maximally specific properties;

[66] Theorem 2 is a slightly strengthened variant of a result in Dietrich and List (2016). While the original result was proved for the equivalents of dilemma-free rightness functions, we drop the "dilemma-free" restriction. As in the case of Theorem 1, we have checked that the result still holds without it.

[67] Strictly speaking, the property $P_{x,K}$ has non-empty extension – as we normally require – only if x is available in context K. Nothing hinges on this.

- for any options x and y and any context K, $\{P_{x,K}\} \geq \{P_{y,K}\}$ if and only if x is deemed right by R in context K while y is available (i.e., $x \in R(K)$ and $y \in [K]$).

So, Axiom 2 is trivially satisfied when all maximally specific properties are admissible as candidates for normatively relevant properties. Yet, the reasons structure which we have just constructed is not very illuminating. It accounts for the given permissibility verdicts simply by stipulating that "being option x in context K" is "better" than "being option y in context K" whenever x is deemed right in context K while y is available. From a substantive perspective, such a representation can be criticized for being gerry-mandered and not identifying any non-*ad-hoc* "right-making" or "good-making" features.

The present considerations show that before we can usefully ask whether a given rightness function has a reason-based representation, we must at least provisionally commit ourselves to some auxiliary hypothesis concerning the set $\mathcal{P}$. In other words, we must take a view on which properties are admissible candidates for right-making or good-making features and which are not.

B.3 How demanding is representability in a teleological format?

Finally, we show that the conditions for reason-based representability with a transitive weighing relation are more demanding than those for reason-based representability *simpliciter*. To make this point, we give an example of a rightness function that admits a reason-based representation relative to a given set $\mathcal{P}$ of admissible properties, but only with an intransitive weighing relation.

Let $X = \{x, y, z\}$. Suppose that, for every non-empty subset Y of X, there is a context K such that $[K] = Y$. The rightness function deems y impermissible whenever option x is available; all other options are always permissible. Formally, $R(K) = [K]$ if x is not in $[K]$, and $R(K) = [K]\backslash\{y\}$ if x is in $[K]$. (Note that $[K]\backslash\{y\} = [K]$ if y is not in $[K]$.) Suppose that $\mathcal{P}$ contains only option properties, where x, y, and z instantiate distinct property sets, say $\mathcal{P}(x) = \{P\}$, $\mathcal{P}(y) = \{Q\}$, and $\mathcal{P}(z) = \{R\}$. This rightness function is representable by a reasons structure $\mathcal{R} = \langle N, \geq \rangle$ in which

- $N(K) = \{P, Q, R\}$ for every context K, and
- $\{P\} > \{Q\}$, $\{P\} \equiv \{R\}$, $\{Q\} \equiv \{R\}$, $\{P\} \equiv \{P\}$, $\{Q\} \equiv \{Q\}$, $\{R\} \equiv \{R\}$.

The weighing relation is clearly intransitive, and it is easy to see that no other reasons structure can represent the same rightness function. Accordingly, the rightness function admits no teleological representation, unless we enrich the set $\mathcal{P}$ of properties we are prepared to invoke.