Choice-theoretic deontic logic

Franz Dietrich[1] & Christian List[2]

July 2017[3]

preliminary draft

Abstract

Rational choice theorists and deontic logicians both study actions, yet using very different approaches and tools. This paper introduces some choice-theoretic concepts – feasible options, choice contexts, choice functions, rankings of options, and reasons structures – into deontic logic. These concepts are used to define a simple 'choice-theoretic' language for deontic logic, and four 'choice-theoretic' semantics for that language, called *basic*, *behavioural*, *ranking-based* and *reason-based* semantics, respectively. We compare these semantics in terms of the strength of their entailment relations, and characterize precisely the 'gaps' in strength between weaker and stronger ones of these semantics.

1 Introduction

Formal decision theory, or *choice theory*, provides us with various models of an agent's choice behaviour, for instance models in terms of a ranking of options (e.g., Samuelson 1948, Sen 1993, Bossert and Suzumura 2009) or in terms of a reasons structure (Dietrich and List 2016). While choice theory is more often interpreted descriptively, as studying *actual* choice, we here

[1] Paris School of Economics & CNRS

[2] London School of Economics

[3] Earlier versions of this work have been presended at *Logics, Decisions and Interactions* (Lorentz Centre, Leiden, October 2017), the seminar of THEMA (University of Cergy-Pontoise, May 2017) and *TARK* (University of Liverpool, July 2017). We thank the audiences at these occasions for helpful feedback.

interpret it purely normatively, as studying *rational* or *moral* choice. So for us a ranking of options will capture normative comparisons, not preferences, and a reasons structure will capture normative reasons, not motivating reasons. The choice-theoretic approach to morality or more broadly normativity has recently received much attention, in terms of both ranking-based models (Brown 2011) and reason-based models (Dietrich and List 2017). Ranking-based models focus on consequentialist and universalist morality, whereas reason-based models also address non-consequentialism and relativism.

Choice theory shares its focus on action with another field: deontic logic, the logic of obligations and permissions.[4] Yet there is a striking disconnect between these fields. Each uses its own tools and concepts. Choice theorists use objects such as options, choice contexts, rankings of options, or reasons structures. These objects are largely absent from standard syntax and semantics of deontic logic. Deontic logicians instead use objects such as possible worlds, modal operators and accessibility relations.

There are of course analogies between choice theory and deontic logic. One might compare *options* in choice theory with *worlds* in deontic logic. One might compare *rankings of options* with *rankings of worlds* – when the latter are introduced in deontic logic, as done occasionally.[5] And one might compare *feasible* options offered by a choice context with *agentially possible* worlds – when agential (or causal) possibility is introduced in deontic logic as a second modality besides moral permissibility.

But these analogies do not go very far and are ultimately problematic. The main problem is that worlds differ considerably from options. Worlds capture *everything*. Yet choice options must be specified coarsely enough to be repeatable in different choice contexts, as choice theorists insist. Otherwise much of choice theory would become trivial or vacuous.[6] Worlds are by

[4] For an introduction to standard modal and deontic logics, see Priest (2001) and Gabbay (2013). Less standard branches or cousins of deontic logic include preference logics (e.g., Hanson 2001, Liu 2010), and logics of action in the form of STIT logics or dynamic logics (for a review see Segerberg et al. 2013).

[5] A ranking of worlds might be used to then define a world as *permissible* if it ranks highest among all worlds, or among those worlds which are agentially possible from the perspective of the actual world. Rankings of works are used in preference logics (e.g., Liu 2010).

[6] Choice made in one context could otherwise never be related to choice in another. Questions of cross-context robustness or consistency of behaviour could thus not even be addressed. More formally, any pattern of choices across contexts could trivially be explained in terms of some ranking of the options, since it is easy to rank the options such

definition unrepeatable: they describe everything, not just the action chosen by the agent, but also the full choice context, and so each world is compatible with only one choice context. The point could be elaborated, but we hope readers have been convinced (if they weren't already) that mainstream deontic logic is largely disconnected from mainstream choice theory in that it contains no direct counterparts of some basic choice-theoretic objects, and vice versa.

Our goal is to show how deontic logic would look like were it directly guided by choice theory. That is, we aim for 'choice-theoretic deontic logic(s)'. We first define a simple propositional language for deontic logic, inspired by choice theory. Some of its atomic sentences, the *choice sentences*, correspond to *options* in choice theory. We then introduce four possible-worlds semantics for this language: *behavioural*, *ranking-based*, *reason-based*, and *basic* semantics, respectively. Each of them will explicitly specify (i) which option the agent *chooses* and (ii) what the *choice context* is. This puts some basic choice-theoretic concepts at nonecentre stage. The four semantics differ in how they capture the deontic modality: here behavioural semantics uses a *choice function* (the most general object by which choice theorists capture behaviour), ranking-based semantics uses a *ranking of options* (not one of worlds), reason-based semantics uses a *reasons structure*, and basic semantics uses an *accessibility relation* between worlds. So the first three semantics capture the deontic modality using choice-theoretic constructs, while the fourth semantics, basic semantics, uses an accessibility relation, making that logic slightly less 'choice-theoretic'.

Different theorems will establish the exact relationships between our four logics. In a nutshell, the behavioural, ranking-based and reason-based logics each strengthen the basic logic, by validating more sentences and more inferences.

In sum, we make the choice-theoretic and logical approaches to normativity mutually comparable by recasting one approach in the framework of the other, i.e., by recasting choice-theoretic models in logical terms. This uncovers the implicit deontic commitments of each choice-theoretic model. Is such a model for instance committed to the principle 'ought implies can'? Where does it agree or disagree with standard deontic logic?

that in any choice context the agent chooses a top-ranking option among the currently feasible options. Indeed it suffices to rank each option which is chosen (in the only context where it is feasible) above all options which are not chosen (in that context).

2 Choice-theoretic models of morality

This section recapitulates the choice-theoretic background of this paper. After some preliminaries (Section 2.1), we define the three choice-theoretic models on which three of our semantics will later be based, respectfully (Sections 2.2-2.4). We shall be brief; for details see Dietrich and List (2017).

2.1 Preliminaries: options and choice contexts

Before defining the three choice-theoretic models, we here introduce what they have in common, i.e., options and contexts.

Let X be a fixed non-empty set: the set of all *options* an agent might ever encounter. A *(choice) context* is a situation in which an agent has to choose between certain options in X. Formally, a context is an object K which comes with a non-empty set of ('*feasible*') options denoted $[K] \subseteq X$ and called the *(choice) menu* of K.

Some dilemma-type context K might offer the menu $[K] = \{k, l\}$ where option k consists in killing a dictator and thereby ending a regime of violent terror, while option l consists in not killing the dictator and thereby leaving the regime in place.

Let $\mathcal{K}$ be a fixed set of contexts: all contexts deemed possible. Different contexts in $\mathcal{K}$ might have overlapping menus, so that the same option belongs to the menu $[K]$ of different contexts K in $\mathcal{K}$.

Choice theorists often identify contexts K with their choice menus $[K]$, by defining a context directly as a menu of feasible options, i.e., a non-empty set of options $K \subseteq X$; the notation $[K]$ is then no longer needed. The context of our kill-or-not example would thus be defined as the set $K = \{k, l\}$ containing the two feasible options. Such a 'thin' individuation of contexts is compatible with our general notion of contexts: it is a special case, obtained by letting $K = [K]$ for all contexts K in $\mathcal{K}$.

But we also allow richer notions of context. Contexts in $\mathcal{K}$ could be pairs $K = (Y, \lambda)$ of a choice menu $Y = [K]$ and an environmental parameter λ describing the environment in which the choice takes place. Our kill-or-not context would then be specified not as $K = \{k, l\}$, but as a pair $K = \{\{k, l\}, \lambda)$, where λ carries some environmental information, say whether there currently is war time or peace time. More generally, λ might capture the *cultural* environment, or the *history* preceding the choice, or information about the agent such as his *identity*, *information* or *awareness*. The point of adding environmental information to contexts is that such information might have moral importance, by affecting which options are permissible. For instance, certain forms of relativism are sensitive to the cultural environment relevant; moral egoism is sensitive to the agent's identity; and whether killing the dictator is right in our example might (under some moral theories) depend on whether there is war time or peace time, as this might affects the status and gravity of killing.

2.2 Representing morality by a rightness function

The most basic and minimal way to model a moral theory consists in specifying a *rightness function*: a function R which to every choice context K in $\mathcal{K}$ assigns a set of feasible options $R(K) \subseteq [K]$, interpreted as the *right* or *permissible* options in context K. Each moral theory has its rightness function. For instance, under classical utilitarianism $R(K) = \{x \in [K] : x$ generates at least as much total happiness as any other feasible option $y \in [K]\}$. In our kill-or-not context, standard consequentialist theories would deem killing right, so that $R(K) = \{k\}$, while some deontological theories which emphasize the asymmetry between action and omission would deem not killing right, so that $R(K) = \{l\}$.

A rightness function R is formally the same object as a *choice function*, except that choice functions by definition never return an empty set while we allow that in some contexts K there is a moral dilemma, i.e., $R(K) = \emptyset$.

A rightness function R fully captures a moral theory's *deontic content*, i.e., the theory's body of permissibility verdicts. By contrast, R fails to capture the reasons or justifications which a theory offers as the basis of its permissibility verdicts. So R captures *what* is right, in various situations, not *why* it is right. We now turn to two other representations of a moral theory, ranking-based and reason-based representations, which each go beyond a theory's deontic content, though in very different ways.

2.3 Representing morality by a ranking of options

Many moral theories can be represented not just by a rightness function, but also by a *ranking* of the options, i.e., a binary relation $\succeq$ on X where $x \succeq y$ is usually taken to mean that option x is *at least as good* as option y. Different theories rank options differently; for instance classical utilitarianism ranks options in terms of total happiness, so that $x \succeq y$ if and only if x generates at least as much total happiness as y.

Any such ranking $\succeq$ implies a particular rightness function R. This rightness function deems a feasible option in a context K to be right just in case it outranks all feasible options:

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

The ranking is said to *explain* R and to be a *(ranking-based) explanation* of R. Not every rightness function R is explicable by a ranking, and even when such an explanation exists it need not be unique. An extensive choice-theoretic literature investigates under which conditions on R there exist ranking-based explanations, sometimes with extra requirements on the ranking such as transitivity.[7]

There are at least three reasons for why many important moral theories are *not* representable by a ranking of the options. First, the theory need not build on any notion of (relative) goodness: it need not have axiological foundations. Second, even if the theory builds on goodness comparisons, a ranking of options need not help since goodness might come from the option-context combination rather than the option alone: the theory might be non-consequentialist. Third and more subtly, even if the theory is consequentialist, it might deem an option x better than another y in *one* context while deeming y better than x in *another* context: the theory might be relativist. This was quick. We shall return to non-consequentialism and relativism in the next section. For now we retain that rankings of options can represent some but not all moral theories; and more formally, they can explain some but not all rightness functions.

2.4 Representing morality by a reasons structure

Many moral theories are representable by a *reasons structure*. A reasons structure specifies 'what matters' and 'how it matters': it specifies (i) which

[7] Giving ranking-based explanations becomes trivially possible (but unilluminating) if one is permitted to re-individuate the options in X (and recast the rightness function accordingly). After sufficiently enriching the options, they will be so specific that each is feasible in just one context; one can then explain choice trivially by ranking each option that is chosen (in the only context where it is feasible) above each option that is not chosen (see Dietrich and List 2017 for technical details). To avoid triviality, we keep X fixed and thus forbid option re-individuation.

properties are normatively relevant in each context, and (ii) how combinations of properties are weighed relative to one other. To make this precise, we fist need the concept of property. Properties are features that options x may or may not have in contexts K where they are feasible. For instance, an option may have the property that someone dies, or that someone is made happy, or that the option is the only feasible one in the context. So properties are features of option-context pairs. Formally, an *option-context* pair is a pair (x, K) where K is a context in $\mathcal{K}$ and x is a feasible option in $[K]$.[8] A *property* is some object p which determines a set of option-context pairs $[p]$, the *extension* of p. When $(x, K) \in [p]$ we say that (x, K) *has* or *satisfies* the property, or that x *has* or *satisfies* the property *in context* K. One might specify a property purely extensionally, by defining p as the set option-context pairs satisfying it; then $p = [p]$. But we also allow an intensional notion of properties. Here properties do not reduce to their extensions, but somehow go beyond. Distinct properties p and q can then be extensionally equivalent ($[p] = [q]$) and yet potentially play different moral roles. For instance the properties p of a maximal number of happy people and q of a minimal number of unhappy people differ only intensionally.

A property p is called

- a *(pure) option property* if its satisfaction does not depend on the context, i.e., if two option-context pairs involving the same option but possibly distinct contexts either both have the property (belong to $[p]$) or both do not have the property. Examples are the property that the option does not involve lying or that it saves a life.

- a *(pure) context property* if its satisfaction does not depend on the option, i.e., if two option-context pairs involving the same context but possibly distinct options either both have the property (belong to $[p]$) or both do not have it. Examples are the property that only one option is feasible, or that there is a feasible option which saves a life, or that the choice happens in a traditional Indian environment.

- a *relational property* if its satisfaction depends on both the option and the context, i.e., if the property is neither an option property nor a context property. An example is the property that the option is the

[8] We build feasibility of the option into the notion of an option-context pair, as a slight departure from Dietrich and List (2017).

only currently feasible one saving a life. Whether this property holds indeed depends both on the option (does the option save a life?) and the context (does the context offer other feasible options that save a life?).

There is an abundance of properties one might imagine. For each set of option-context pairs one might imagine a property whose extension is that set. Most of these properties are artificial and cannot plausibly play a moral role. We thus fix a set $\mathcal{P}$ of properties deemed to be permissible candidates for playing a moral role (and specified extensionally or intensionally). Let

- $\mathcal{P}(x, K)$ be the set of all properties in $\mathcal{P}$ satisfied by the option-context pair (x, K),
- $\mathcal{P}(x)$ be the set of all *option* properties in $\mathcal{P}$ satisfied by the option x (independently of the context),
- $\mathcal{P}(K)$ be the set of all *context* properties in $\mathcal{P}$ satisfied by the context K (independently of the option).

A *reasons structure* is defined to be a pair $\mathcal{R} = (N, \geq)$ of two objects:

- a function N, the *normative relevance function*, which for any context $K \in \mathcal{K}$ specifies a set $N(K) \subseteq \mathcal{P}$ of properties, the *normatively relevant properties in context* K. We place a single requirement on this function. Informally, changes in what is normatively relevant must stem from changes in context properties (i.e., should not be 'arbitrary'). Formally, whenever two contexts $K, K' \in \mathcal{K}$ have identical context properties they induce identical normatively relevant properties: $\mathcal{P}(K) = \mathcal{P}(K') \Rightarrow N(K) = N(K')$.
- a binary relation $\geq$ on the set of property combinations (subsets of $\mathcal{P}$), the *weighing relation*. We interpret $S \geq T$ to mean that the property combination S weighs normatively at least as much as ('*outweighs*') the property combination T.

Given such a reasons structure, each feasible option x in a context K has a *moral description*, defined as the set $N(x, K)$ of all properties which pertain to x in context K *and* are normatively relevant in context K. So $N(x, K) = \mathcal{P}(x, K) \cap N(K)$.

Several moral theories are representable by reasons structures (see Dietrich and List 2017). Consider for instance classical utilitarianism. Here the set $N(K)$ contains all *happiness properties*, i.e., all properties of type "the option generates total happiness h", denoted p_h, where h ranges over a fixed set of possible happiness levels (a fixed subset of $\mathbb{R}$, e.g., $[0, \infty)$). Each option x generates some happiness level h_x, so that its moral description in any context K is $N(x, K) = \{p_{h_x}\}$. Further, the utilitarian weighing relation satisfies $p_h \geq p_{h'}$ if and only if $h \geq h'$ (for any two happiness levels h and h').

This utilitarian reasons structure $(N, \geq)$ is special in three ways: (i) only option properties are ever relevant, i.e., belong to $N(K)$ for some context K; (ii) there are never any changes in what is relevant, i.e., $N(K)$ is the same for all contexts K; (iii) any option is morally described by a single property, i.e., $N(x, K)$ is singleton for all option-context pairs (x, K). We refer to these three conditions as consequentialism, universalism, and monism, respectively. In general, a reasons structure $(N, \geq)$ is:

- *consequentialist* or *context-unrelated* if all $N(K)$ $(K \in \mathcal{K})$ contain only option properties, and *non-consequentialist* or *context-related* otherwise;

- *universalist* or *context-invariant* if all $N(K)$ $(K \in \mathcal{K})$ are the same, and *relativist* or *context-variant* otherwise;

- *monistic* if $N(x, K)$ contains a single property for all option-context pairs (x, K), and *pluralistic* otherwise.

These conditions can be combined at will; an example is consequentialism together with relativism and pluralism. Most if not all $2^3 = 8$ combinations are prima facie plausible and have their counterparts in moral philosophy. We again refer to Dietrich and List (2017) for details.

Just as a ranking of options, so a reasons structure $\mathcal{R} = (N, \geq)$ implies a particular rightness function R, of which we say that it *explains* $\mathcal{R}$ or is a *(reason-based) explanation* of $\mathcal{R}$. This rightness function specifies for any context $K \in \mathcal{K}$ that a feasible option is right just in case its normatively relevant properties outweigh those of each feasible option:

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

The same rightness function R can have different reason-based explanations – or *no* reason-based explanation, although this case does normally not occur if R is plausible and $\mathcal{P}$ includes all important properties. The main reason why reason-based explanations usually exist is that they can explicitly accommodate non-consequentialist and relativist theories, unlike ranking-based explanations. The question of when exactly reason-based explanations exist has here again a clear-cut answer in the form of necessary and sufficient conditions on the rightness function (Dietrich and List 2017). Reason-based explanations 'explain' rightness in a richer and more proper sense than ranking-based explanations, by giving a substantive account of justification of choice. In a sense, a reasons structure captures a moral theory, while a ranking of options captures not much more than a theory's deontic content.

3 A choice-theoretic language of deontic logic

The propositional language to be defined enriches the standard deontic language in three ways. First, we use two distinct types of atomic sentences: *choice sentences* interpreted as saying that a certain option is chosen, and *basic descriptive sentences* which could be interpreted as describing properties of the choice and/or the context. Second, we add the *agential* modality of what is possible or necessary through the agent's choice, in addition to the *deontic* modality of what is permissible or obligatory. So we can express that the agent *can* help his noneneighbour, and that he is morally *obliged* to do so. The choice sentences as well as the agential modality make our language 'choice-theoretic', as they correspond to two choice-theoretic concepts, *options* and *feasibility.* Third, we add a strict conditional, to express that a conclusion holds in *all worlds* in which a premises hold. Having a strict conditional as well as both modalities in the language gives us key resources for capturing moral discourse. This can be illustrated with the notorious principle 'ought implies can'. This principle can be expressed by a schema of sentences of the form 'if it is obligatory that ϕ, then it is possible that ϕ' (where ϕ is any sentence). Such a sentence contains the deontic modality of obligation, the agential modality of possibility, and the strict conditional. Using instead a material conditional would have been inadequate, since 'ought implies can' is meant to express a law-like principle rather than a contingent fact about the actual world.[9]

Formally, let X and $\mathcal{P}$ be two (disjoint) sets of atomic sentences, called

[9] Many other moral laws, such as 'If someone helps you, you ought to thank him', would be inexpressible if we only had the material conditional at our disposal.

choice sentences and *basic descriptive* sentences. Note that we use the same symbols as earlier for the set of options and the set of properties. This is explained by the following notational convention.

Convention: Choice sentences and options are denoted by the same symbols, i.e., each x in X stands either for the sentence that a certain option is chosen, or for the option itself. Later when introducing reason-based semantics we will also convene that each p in $\mathcal{P}$ can stand not only for a basic descriptive sentence, but also for a corresponding property.

Our language, denoted $\mathcal{L}$, contains all sentences constructible from the atomic sentences in X and $\mathcal{P}$ using the operators of negation $\neg$, conjunction $\wedge$, obligation $\mathbf{O}$, agential necessity $\Box$, and strict conditional$\Rightarrow$. Formally, $\mathcal{L}$ is the (smallest) set which contains all sentences in X or $\mathcal{P}$ and is closed under construction, i.e., whenever $\mathcal{L}$ contains ϕ and ψ, then $\mathcal{L}$ contains $\neg\phi$ (*not* ϕ), $(\phi \wedge \psi)$ (ϕ *and* ψ), $\mathbf{O}\phi$ (*it is obligatory that* ϕ, in short *obligatorily* ϕ), $\Box\phi$ (*it is agentially necessary that* ϕ, in short *unpreventably* ϕ), and $(\phi \Rightarrow \psi)$ (*if* ϕ *then* ψ). So $\mathcal{L}$ contains sentences such as x (e.g., 'you kill Mr X'), p (e.g., 'Mr X threatens your family's life'), q (e.g., 'there is an escape route'), $(p \wedge \neg q) \Rightarrow \neg\mathbf{O}\neg x$ (e.g., 'whenever Mr X threatens your family's life and there is no escape route, then you are not obliged not to kill Mr X'), and so on.

Other truthfunctional operators such as disjunction $\vee$, material implication $\rightarrow$ and material bi-implication $\leftrightarrow$ are definable in the usual way from $\neg$ and $\wedge$: $(\phi \vee \psi)$ stands for $\neg(\neg\phi \wedge \neg\psi)$, $(\phi \rightarrow \psi)$ stands for $\neg(\phi \wedge \neg\psi)$, and so on. Moreover, we introduce the duals of the modal operators $\mathbf{O}$ and $\Box$, namely permission $\mathbf{P}$ and agential possibility $\diamond$. Formally, $\mathbf{P}\phi$ stands for $\neg\mathbf{O}\neg\phi$ and reads *it is permissible that* ϕ, in short *permissibly* ϕ; and $\diamond\phi$ stands for $\neg\Box\neg\phi$ and reads *it is agentially possible that* ϕ, in short *feasibly* ϕ.

We finally define the *universal* modality, given by an operator $\blacksquare$ of conceptual or logical necessity, and its dual possibility operator $\blacktriangleleft$. Formally, $\blacksquare\phi$ stands for $\tau \Rightarrow \phi$ where τ is a truth-functional tautology (e.g., $(\phi \vee \neg\phi)$) and reads *it is conceptually necessary that* ϕ, in short *always* ϕ; and $\blacktriangleleft\ \phi$ stands for $\neg\blacksquare\neg\phi$ and reads *it is conceptually possible that* ϕ, in short *sometimes* ϕ We could equivalently have taken conceptual necessity $\blacksquare$ as a primitive operator of the language and defined the strict conditional $\phi \Rightarrow \psi$ as $\blacksquare(\phi \rightarrow \psi)$. The universal modality and the strict conditional are interdefinable.

4 Four semantics for the language

We now define four possible-worlds semantics for our language. They will construct their worlds from options and contexts, and base their deontic modality on either a rightness function, or a ranking of options, or a reasons structure, or the more standard semantic tool of an accessibility relation between worlds. These semantics are *choice-theoretic* in two senses, namely firstly through their notion of worlds as option-context pairs, and secondly (except for the fourth-mentioned semantics) through their deontic modality.

We first introduce what is common between the four semantics (Section 4.1), and then define the four semantics (Section 4.2). Logical relations between the semantics are addressed later in Section 5.

4.1 Commonalities between the four semantics

Worlds as option-context pairs: For each of our four semantics – the behavioural, ranking-based, reason-based, and basic semantics – an interpretation is a triple $(W, v, *)$ in which W is a set of worlds, v is a truth function on $\mathcal{P}$, and $*$ is some further object, which is a rightness function or ranking or reasons structure or accessibility relation, depending on the type of semantics. The worlds in W will not be primitive objects, but option-context pairs. To construct a set of worlds W, one first chooses a set $\mathcal{K}$ of contexts, where by Section 2 a context is an object K which determines a choice menu $[K] \subseteq X$. The set of contexts then gives rise to a set of option-context pairs (worlds) given by $W = \{(x, K) : K \in \mathcal{K}, x \in [K]\}$.

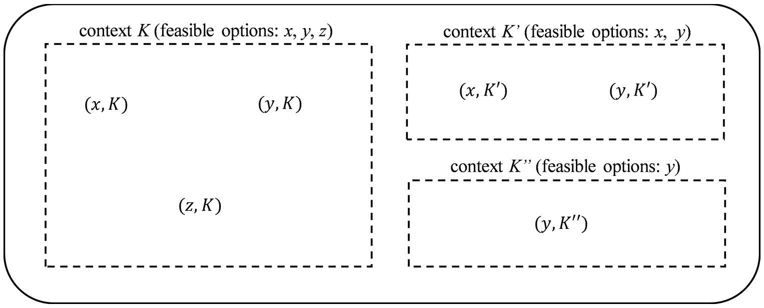
context K (feasible options: x, y, z)
(x,K)
(y,K)
(z,K)
context K' (feasible options: x, y)
(x,K')
(y,K')
context K'' (feasible options: y)
(y,K'')

Figure 1: A set of worlds W with 3 contexts, 3 options, and 6 worlds

Figure 1 gives an example with 3 contexts, 6 worlds, and three options x, y, z in X. The simplest way to construct a W is to fix a set $\mathcal{K}$ of choice menus $K \subseteq X$, thereby identifying contexts with choice menus, so that $W = \{(x, K) : x \in K \in \mathcal{K}\}$. If we for instance put $\mathcal{K} = \{\{x, y\}, \{x, y, z\}\}$, we obtain 5 worlds: $W = \{(x, \{x, y\}), (y, \{x, y\}), (x, \{x, y, z\}), (y, \{x, y, z\}), (z, \{x, y, z\})\}$. Another way to construct W is to define contexts as pairs $K = (Y, \lambda)$ of a choice menu $Y = [K]$ taken from a certain set $\mathcal{Y}$ of possible choice menus and an environmental parameter λ taken from a certain set Γ of possible environmental parameters. Here the set of worlds becomes $W = \{(x, (Y, \lambda)) : x \in Y \in \mathcal{Y}, \lambda \in \Gamma\}$.

Convention: Whenever we invoke a set of worlds W, its underlying set of contexts will be denoted $\mathcal{K}$. For any world w in our sense (an option-context pair) we denote its option by x_w and its context by K_w; so $w = (x_w, K_w)$.

Agential possibility: Our notion of worlds as option-context pairs induces a notion of agential possibility. Note that, intuitively, a world is agentially possible at the actual world if the agent *could* bring it about at the actual world. Now the agent can control the chosen option, but not the context: he chooses the option, not the context. So, given a set of worlds W, we define a world $w' \in W$ to be *(agentially) possible* at a world $w \in W$ if the context is the same: $K_{w'} = K_w$. If w' is possible at w, then the choice made at w' is automatically feasible at w: $x_{w'} \in [K_w]$.

Interdefinability of W and $\mathcal{K}$: Although one naturally thinks of the set of worlds W as constructed from the set of contexts $\mathcal{K}$, the two sets are in fact interdefinable. To see why, let us start not from a set of contexts $\mathcal{K}$, but from any set W, now defined as any set consisting of certain pairs (x, K) of an option x in X and *some* (so far arbitrary) object K. We can then retrieve the implicit set of contexts as being $\mathcal{K} = \{K : (x, K) \in W$ for some $x \in X\}$, where the choice menu offered by a context $K \in \mathcal{K}$ is $[K] = \{x \in X : (x, K) \in W\}$. The interdefinability of W and $\mathcal{K}$ implies that we could define interpretations either as triples $(W, v, *)$ or as triples $(\mathcal{K}, v, *)$. We choose the former, to maximize familiarity to readers used to possible-worlds semantics.

4.2 The four semantics defined

Interpretations. We define a *basic*, *behavioural*, *ranking-based*, or *reason-based interpretation* of the language $\mathcal{L}$ as, respectively, a triple (W, v, ρ), (W, v, R), $(W, v, \succeq)$, or $(W, v, \mathcal{R})$, in which

- W is the set $\{(x, K) : K \in \mathcal{K}, x \in [K[\}$ of all option-context pairs (*worlds*) given by some set of contexts $\mathcal{K}$,
- v is a function from $W \times \mathcal{P}$ to $\{T, F\}$, the *truth function*, mapping any pair (w, p) of a world and a basic descriptive sentence to the truth value $v_w(p)$ of p at the world w,

and, respectively,

- ρ is a binary '*accessibility*' relation on W ($w\rho w'$ means that world w' is permissible at world w), where ρ respects agential possibility in that each world accesses only worlds which are agentially possible, i.e., have same context,
- R is a rightness function on $\mathcal{K}$ ($R(K)$ contains the right options in context K),
- $\succeq$ is a binary '*ranking*' relation on X ($x \succeq y$ means that x is at least as good as y).
- $\mathcal{R}$ is a reasons structure relative to the set of options, the set of contexts $\mathcal{K}$, and the set $\mathcal{P}$ re-interpreted as containing properties.

Re-interpreting sentences in $\mathcal{P}$ as properties of option-context pairs is very natural, since option-context pairs are simply worlds, and to a sentence corresponds the property of being true at a world. Formally:

Convention (for reason-based semantics): Given a set of words W and a truth function v, we identify each basic descriptive sentence p in $\mathcal{P}$ with the equally-labelled property *that it is true*, satisfied by those option-context pairs (worlds) $w = (x, K) \in W$ for which $v_w(p) = T$.

Truth at a world of an interpretation. Let $\mathcal{M}$ be an interpretation of any of the four kinds, i.e., a basic one (W, v, ρ) or behavioural one (W, v, R) or ranking-based one $(W, v, \succeq)$ or reason-based one $(W, v, \mathcal{R})$. Truth of an arbitrary sentence ϕ in $\mathcal{L}$ at a world w of $\mathcal{M}$ is denoted $\mathcal{M}, w \vDash \phi$ and defined via the following recursive instructions:

- $\mathcal{M}, w \vDash x$ (where $x \in X$) if and only if x is chosen at w, i.e., $x = x_w$,
- $\mathcal{M}, w \vDash p$ (where $p \in \mathcal{P}$) if and only if $v_w(p) = T$,
- $\mathcal{M}, w \vDash \neg\phi$ if and only if $\mathcal{M}, w \nvDash \phi$, i.e., not $\mathcal{M}, w \vDash \phi$,
- $\mathcal{M}, w \vDash (\phi \wedge \psi)$ if and only if $\mathcal{M}, w \vDash \phi$ and $\mathcal{M}, w \vDash \psi$,
- $\mathcal{M}, w \vDash \Box\phi$ if and only if ϕ holds in all agentially possible worlds, i.e., $\mathcal{M}, w' \vDash \phi$ for all world $w' \in W$ sharing the context with w,
- $\mathcal{M}, w \vDash (\phi \Rightarrow \psi)$ if and only if ψ holds in all worlds where ϕ holds, i.e., for all worlds w' in W, whenever $\mathcal{M}, w' \vDash \phi$ then $\mathcal{M}, w' \vDash \psi$ (equivalently, $\mathcal{M}, w' \vDash (\phi \rightarrow \psi)$),
- $\mathcal{M}, w \vDash \mathbf{O}\phi$ if and only if, respectively,
  - ϕ holds in all accessible worlds, i.e., $\mathcal{M}, w' \vDash \phi$ for all worlds $w' \in W$ such that $w\rho w'$,
  - ϕ holds in all agentially possible world in which the chosen option is in $R(K)$, i.e., $\mathcal{M}, w' \vDash \phi$ for all worlds $w' \in W$ with the same context K as in w and with some option in $R(K)$,
  - ϕ holds in all agentially possible world in which the chosen option ranks top, i.e., $\mathcal{M}, w' \vDash \phi$ for all worlds $w' \in W$ with the same context K as in w and with an option x such that $x \succeq y$ for each y in $[K]$,
  - ϕ holds in all agentially possible worlds whose normatively relevant properties rank top, i.e., $\mathcal{M}, w' \vDash \phi$ for all worlds $w' \in W$ with the same context K as in w and with an option x satisfying $N(x, K) \geq N(y, K)$ for all $y \in [K]$ (i.e., satisfying $N(w') \geq N(w'')$ for all worlds $w'' \in W$ with context K).

The first two instructions settle the truth values of atomic sentences of the two types. The other instructions give the truth conditions of each operator in the language. The truth condition for the deontic operator **O** marks the (only) difference between the four semantics; it is based on ρ, R, $\succeq$, or $\mathcal{R}$.

5 How are the four semantics related?

How do the four semantics compare to one another in terms of strength? For instance, does basic semantics validate more or fewer sentences and entailments than behavioural semantics? We first clarify the hierarchy of logical strength between the four semantics (Section 5.1), then characterize the precise "logical gap" between weaker and stronger of the four semantics (Section 5.2), and finally consider some refinements of our four semantics obtained by placing requirements on interpretations such as transitivity of the ranking or consequentialism of the reasons structures (Section 5.3).

5.1 The hierarchy of logical strength

As it turns out, behavioural interpretations are equivalent to particular basic interpretations, and ranking- and reason-based interpretations are equivalent to particular behavioural interpretations. Remarks 1-3 state this formally. We call one interpretation $\mathcal{M}$ (of any of our four types) *equivalent* to another one $\mathcal{M}'$ (of the same or another of our types) if $\mathcal{M}$ and $\mathcal{M}'$ have the same set of worlds W and make the same sentences true at each world in W; in particular they make the same *atomic sentences in* $\mathcal{P}$ true at each world, so have the same truth function v on $\mathcal{P}$.

Remark 1. Every behavioural interpretation (W, v, R) is equivalent to a basic one (W, v, ρ), by taking $w \rho w'$ to mean that w' has the same context K as w and some option from $R(K)$.

Remark 2. Every ranking-based interpretation $(W, v, \succeq)$ is equivalent to a behavioural one (W, v, R), obtained by letting R be the rightness function implied (explained) by $\succeq$.

Remark 3. Every reason-based interpretation $(W, v, \mathcal{R})$ is equivalent to a behavioural one (W, v, R), obtained by letting R be the rightness function implied (explained) by $\mathcal{R}$.

Remarks 1-3 imply the hierarchy shown in Figure 2: basic semantics (denoted $BASIC$) is strengthened by behavioural semantics (denoted $BEHA$), which is in turn strengthened by ranking-based semantics (denoted $RANK$) and reason-based semantics (denoted $REAS$). As usual, one semantics is 'stronger' than or 'strengthens' another if it yields at least the same logical entailments among sentences, hence has at least the same logically valid sentences. To put this formally, note first that any *(possible-worlds) semantics*

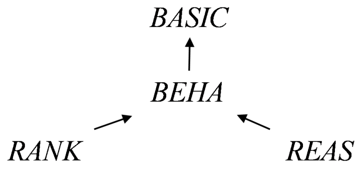

Figure 2: Relationship between the four semantics (arrows point from a stronger to a weaker semantics)

for $\mathcal{L}$ defines a class $\mathcal{S}$ of admissible interpretations for $\mathcal{L}$ (for instance the class of behavioural interpretations in case of $BEHA$). It further defines an *entailment relation* $\models$, where a set of sentences $\Phi \subseteq \mathcal{L}$ *entails* $\phi \in \mathcal{L}$ – written $\Phi \models \phi$ – if truth of all sentences in Φ at a world of an interpretation in $\mathcal{S}$ implies truth of ϕ at this world of the interpretation. A sentence ϕ entailed by the empty set is *logically valid*; we then write $\models \phi$ rather than $\emptyset \models \phi$. One semantics with entailment relation $\models$ *strengthens* another with entailment relation $\models'$ if, for all $\Phi \subseteq \mathcal{L}$ and $\phi \in \mathcal{L}$, $\Phi \models' \phi$ implies $\Phi \models \phi$ (hence in particular$\models \phi$ implies $\models' \phi$). Two semantics are *equivalent* if they yield the same entailment relation (hence the same logical truths).

5.2 The gap from one semantics to another

But *how* does $BEHA$ strengthen $BASIC$? And *how* do $RANK$ and $REAS$ strengthen $BEHA$? Theorems 4-6 will answer these questions by characterizing the logical gap between $BEHA$ and $BASIC$, between $RANK$ and $BEHA$, and between $REAS$ and $BEHA$. In each theorem, the logical gap in question will be characterized both semantically and syntactically. For instance the gap between $BEHA$ and $BASIC$ will be characterized

- *semantically* by specifying a condition on the accessibility relation of a basic interpretation under which the interpretation becomes equivalent to a behavioural interpretation,

- *syntactically* by specifying a schema of sentences which are logically true under $BEHA$ but not under $BASIC$.

These results can guide the search for adequate semantics for deontic logic, because the question of whether to strengthen $BASIC$ to $BEHA$, and further to $RANK$ or rather to $REAS$, becomes the question of whether to endorse the additional semantic or syntactic features coming with that strengthening.

We begin with the gap between $BASIC$ and $BEHA$. *Semantically*, this gap is closed by the following condition on a basic interpretation (W, v, ρ) or more precisely on its accessibility relation ρ:

Ai (act-independence): If two worlds w and w' share the same context K, then they access the same worlds, i.e., for all worlds $\bar{w} \in W$, $w\rho\bar{w} \Leftrightarrow w\rho\bar{w}$.

Informally, the agent's action (which constitutes the only different between w and w') does not affect what is permissible. In short, we are not the authors (by our actions) of what is right or wrong. This is very plausible. Note that the interpretation defined in Remark 1 indeed satisfies Ai. The following schema of sentences (axioms) reflects the same idea:

$$Ax1: \ \mathbf{O}\phi \Rightarrow \Box\mathbf{O}\phi, \text{ for } \phi \in \mathcal{L}.$$

Informally, the sentence $\mathbf{O}\phi \Rightarrow \Box\mathbf{O}\phi$ means that whenever something (ϕ) is obligatory, then it is obligatory independently of the action, i.e., the obligation is agentially necessary.

The first theorem states that the condition Ai or the schema $Ax1$ define precisely the gap from $BASIC$ to $BEHA$, in the sense that adding Ai or $Ax1$ makes $BASIC$ equivalent to $BEHA$.

Convention: For any semantics SEM (e.g., $BASIC$),

- if the class of interpretations is restricted by imposing a condition or list of conditions Con on interpretations (e.g., condition Ai), the resulting stronger semantics is denoted $SEM + Con$ (e.g., $BASIC + Ai$),
- if the class of interpretations is restricted by imposing validity at all worlds of a schema of sentences or list of schemas Ax (e.g., schema $Ax1$), the resulting stronger semantics is denoted $SEM + Ax$ (e.g., $BASIC + Ax1$).

Theorem 4. *$BEHA$ is equivalent to $BASIC + Ai$, and to $BASIC + Ax1$.*

It seems highly plausible to impose Ai or $Ax1$. So, someone who endorses $BASIC$ has reason to even endorse $BEHA$.

Next, what is the gap from $BEHA$ to $RANK$? Given a rightness function R, we say that option x *sometimes beats* option y if x is right in at least one context where y is feasible, i.e., there is at least one context K in $\mathcal{K}$ such that $x \in R(K)$ and $y \in [K]$. The concept that an option 'sometimes beats' another corresponds precisely to the classic concept of *revealed preference* in choice theory. Choice theorists say that an option is revealed-preferred to another if sometimes the first option is chosen while the second is feasible. The following condition on a behavioural interpretation (W, v, R), or more precisely on its rightness function R, is then the analogue of a classic choice-theoretic condition due to Richter (1971):

Rev (Richter's *revelation coherence*): If an option x is feasible in a context K and each option feasible in K is sometimes beaten by x, then x is a right in the context, i.e., $x \in R(K)$.

For any sentences ϕ and ψ, let $SB(\phi, \psi)$ denote the sentence$\blacktriangleleft$ $(\mathbf{P}\phi \wedge \diamond\psi)$ expressing that ϕ sometimes beats ψ, or more precisely that sometimes ϕ is permissible while ψ is feasible. The following schema of sentences corresponds to condition *Rev*. It assumes finite X (so that the conjunction indexed by X is well-defined):

$$Ax2: \; (\diamond x \wedge [\wedge_{y \in X}(\diamond y \to SB(x, y))]) \Rightarrow \mathbf{P}x, \text{ for } x \in X.$$

Theorem 5. *$RANK$ is equivalent to $BEHA + Rev$, and (if X is finite) to $BEHA + Ax2$.*

Finally, what is the gap between $BEHA$ and $REAS$? Consider the following two conditions on a behavioural interpretation (W, v, R), or more precisely on its rightness function R. They are obviously satisfied if R is generated by a reasons structure.[10] The first condition says that in any context K the rightness of options depends only on the properties in $\mathcal{P}$ of these options:

$Re1$: For all contexts K and feasible options $x, y \in [K]$, if $\mathcal{P}(x, K) = \mathcal{P}(y, K)$ then $x \in R(K) \Leftrightarrow y \in R(K)$.

The second condition is an analogue of *Rev*, except that options in X are replaced by sets of properties in $\mathcal{P}$. We first adapt our terminology to such

[10] In fact, they are necessary and sufficient for R to be reason-based representable (see Dietrich and List 2016).

sets. A set $S \subseteq \mathcal{P}$ is *feasible* in context K if $S = \mathcal{P}(x, K)$ for some feasible option $x \in [K]$, and *right* in context K if $S = \mathcal{P}(x, K)$ for some right option $x \in R(K)$. A set $S \subseteq \mathcal{P}$ *sometimes beats* $S' \subseteq \mathcal{P}$ if in some context S is right while S' is feasible. This is the condition:

$Re2$: If a set $S \subseteq \mathcal{P}$ is feasible in a context K and each set $S' \subseteq \mathcal{P}$ feasible in K is sometimes beaten by S, then S is right in K.

To state syntactic counterparts of $Re1$ and $Re2$, we write $\widehat{\mathcal{P}}$ for the set of sentences expressing complete states with respect to $\mathcal{P}$, i.e., sentences expressing which sentences in $\mathcal{P}$ hold (those in a particular set $S \subseteq \mathcal{P}$) and which do not (those in the complement $\mathcal{P} \backslash S$). Formally,

$$\widehat{\mathcal{P}} = \{(\wedge_{p \in S} p) \wedge (\wedge_{p \in \mathcal{P} \backslash S} \neg p) : S \subseteq \mathcal{P}\}.$$

Note that $\widehat{\mathcal{P}}$ (and hence, the schemas $Ax3$ and $Ax4$ below) are only defined if $\mathcal{P}$ is finite, so that the conjunctions indexed by S and by $\mathcal{P} \backslash S$ are well-defined. Consider the following schema of sentences:

$$Ax3 : \ (\mathbf{P}p \wedge \diamond(p \wedge x)) \Rightarrow \mathbf{P}x, \text{ for } x \in X \text{ and } p \in \widehat{\mathcal{P}}.$$

Informally, $(\mathbf{P}p \wedge \diamond(p \wedge x)) \Rightarrow \mathbf{P}x$ expresses that whenever p is permissible and can hold together with action x, then that action is permissible. The rationale is that whether an action is permissible is fully determined by the truth values of the sentences in $\mathcal{P}$; so whenever an action x is possible together with a *permissible* combination in $\widehat{\mathcal{P}}$, then x is itself permissible. As shown in the appendix, $Ax3$ is the exact syntactic counterpart of condition $Re1$. Finally, the following schema is the counterpart of condition $Re2$. just as the schema $Ax2$ is the counterpart of condition Rev:

$$Ax4 : \ (\diamond p \wedge [\wedge_{q \in \widehat{\mathcal{P}}} (\diamond q \rightarrow SB(p, q))]) \Rightarrow \mathbf{P}p, \text{ for } p \in \widehat{\mathcal{P}}.$$

Theorem 6. *REAS is equivalent to* $BEHA + Re1, Re2$*, and (if* $\mathcal{P}$ *is finite) to* $BEHA + Ax3, Ax4$*.*

5.3 Refining the semantics by adding conditions on interpretations

Each of our semantics can be strengthened by restricting its set of admissible interpretations. We have already considered some strengthenings: that of $BASIC$ to $BASIC + Ai$, and those of $BEHA$ to $BEHA + Rev$ and $BEHA + Re2, Re3$. There the purpose was to reduce one semantics to another (see Theorems 4-6). Setting that purpose aside, we now mention further strengthenings of interest, i.e., other conditions one might impose on interpretations.

Basic semantics might be strengthened by imposing the following classic condition on the accessibility relation ρ of a basic interpretations (W, v, ρ):

$Dilfree$ (dilemma-freeness): at each world w there is a permissible world, i.e., a world w' such that $w\rho w'$.

Behavioural semantics might once again be strengthened by imposing dilemma-freeness, this time expressed as a condition on the rightness function R of a behavioural interpretation (W, v, R) (we use the same label '$Dilfree$'):

$Dilfree$ (dilemma-freeness): In each context K, at least one feasible option is permissible, i.e., $[K] \neq \emptyset$.

Ranking-based semantics might be strengthened by imposing one or more of the following conditions on the ranking $\geq$ of a ranking-based interpretation $(W, v, \geq)$. The first condition is again dilemma-freeness, now in a ranking-based rendition (yet denoted by the same symbol $Dilfree$):

$Dilfree$ (dilemma-freeness): In each context K, there is a top-ranking feasible option, i.e., an $x \in [K]$ such that $x \succeq y$ for all $y \in [K]$.

Com (commensurability) $\succeq$ is complete, i.e., for all $x, y \in X$, $x \succeq y$ or $y \succeq x$.

$Tran$ (transitivity or teleology): $\succeq$ is transitive, i.e., for all $x, y, z \in X$, if $x \succeq y$ and $y \succeq z$, then $x \succeq z$.[11]

Reason-based semantics might be strengthened in several interesting ways, by imposing one or more of the following conditions on the reasons structure $\mathcal{R} = (N, \geq)$ of a reason-based interpretation $(W, v, \mathcal{R})$. While the first three conditions are reason-based analogues of the above conditions on behavioural interpretations (and shall be denoted by the same symbols), the last three conditions are very different and illustrate the richness of reason-based semantics:

[11] The interpretation of transitivity as teleology follows Broome (2004) and Dietrich and List (2017). One might also require reflexivity as part of teleology.

$Dilfree$ (dilemma-freeness): In each context K, there is a top-ranking feasible option, i.e., an $x \in [K]$ such that $N(x, K) \geq N(y, K)$ for all $y \in [K]$.

Com (commensurability): $\geq$ is complete, i.e., for all property sets $S, S' \subseteq \mathcal{P}$, $S \geq S'$ or $S' \geq S$.[12]

$Tran$ (transitivity or teleology): $\geq$ is transitive, i.e., for all property sets $S, S', S'' \subseteq \mathcal{P}$, if $S \geq S'$ and $S' \geq S''$, then $S \geq S''$.[13]

$Cons$ (consequentialism): Only context-unrelated properties are ever normatively relevant, i.e., for all contexts K all normatively relevant properties in $N(K)$ are context-unrelated.

$Univ$ (universalism): There are no changes in normatively relevant properties, i.e., the set of normatively relevant properties $N(K)$ is the same for all contexts $K \in \mathcal{K}$.

$Univ^*$ (full-relevance universalism): Always all properties are relevant, i.e., $N(K) = \mathcal{P}$ for all contexts K. (This implies $Univ$.)

$Cons^*$ (full-relevance consequentialism): Always all option properties (and no other properties) are relevant, i.e., $N(K)$ consists of all option properties for all contexts K. (This implies $Cons$ and $Univ$.)

$Moni$ (monism): Each option x in any context K has a single normatively relevant property, i.e., $N(x, K)$ is singleton.

Figure 3 refines Figure 2 by adding to it some of our refined reason-based semantics. We conjecture (but have yet to verify) that the graph in 3 stays correct if we we do one or both of the following modifications:

- replacing each semantics by its dilemma-free version, by always imposing $Dilfree$ everywhere,

[12] One might require completeness of $\geq$ only among those pairs of property sets S and S' for which comparisons matter, in the sense that there is a context in which S and S' are the sets of normatively relevant properties of some feasible options x and x', respectively.

[13] One might require transitivity only where comparisons matter, i.e., one might quantify only over those triples of property sets S, S' and S'' such that in at least one context S, S' and S'' are instantiated as the sets of normatively relevant properties of some feasible options x, x' and x'', respectively.

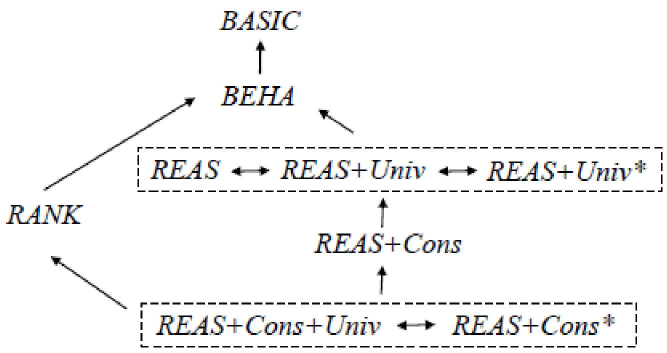

Figure 3: Relationship between some refined semantics (arrows point from a stronger to a weaker semantics, bidirectional arrows indicate equivalence)

- replacing each semantics except $BASIC$ by its commensurable and/or transitive version, by imposing Com and/or $Tran$ everywhere.

In Figure 3, why is $REAS$ equivalent to $REAS+Univ$ and $REAS+Univ^*$, rather than strictly weaker? Why is $REAS + Cons + Univ$ equivalent to $REAS+Cons^*$, rather than strictly weaker? And why does $REAS+Cons+Univ$ strengthen $RANK$? The answers are given by the next three remarks, respectively.

Definition 7. One reasons structure (or ranking of X) is **deontically equivalent** to another one if both imply the same rightness function.

Remark 8. Every interpretation of $REAS$ is equivalent to one of $REAS + Univ$, and even to one of $REAS + Univ^*$. This is because each reasons structure $\mathcal{R} = (N, \geq)$ is deontically equivalent to one that is universalist and if wished even full-relevance universalist (see Dietrich and List 2016, 2017).

Remark 9. Every interpretation of $REAS + Cons + Univ$ is equivalent to one of $REAS+Cons^*$. This is because each consequentialist and universalist reasons structure $\mathcal{R} = (N, \geq)$ is deontically equivalent to a reasons structure that is full-relevance consequentialist.[14]

[14] This new reasons structure $(N', \geq')$ is defined as follows. By universalism we can identify N and N' with fixed sets of properties. Let N' be the set of *all* option properties (so we add to N all missing option properties), and let $S \geq' T$ mean that $S \cap N \geq T \cap N$ (so $\geq'$ pays no attention to the added properties).

Remark 10. Every interpretation of $REAS + Cons + Univ$ is equivalent to one of $RANK$. This is because each consequentialist and universalist reasons structure $\mathcal{R} = (N, \geq)$ is deontically equivalent to some ranking $\succeq$ of X.[15]

Remark 8 has an important implication: when working with $REAS$ – e.g., when checking validity of entailments, or consistency of sets of sentences – we can restrict attention to reason-based interpretations $(W, v, \mathcal{R})$ which are full-relevance universalist. This simplifies the exercise significantly, since the reasons structure $\mathcal{R} = (N, \geq)$ is then given by just one parameter, the weighing relation $\geq$, while N is fixed and trivial. Under full-relevance universalism, we can abbreviate the reasons structure by $\geq$ and the interpretation by $(W, v, \geq)$. Such a special reason-based interpretation $(W, v, \geq)$ is an object as simple as a ranking-based interpretation $(W, v, \succeq)$; the sole difference is that $\geq$ ranks property sets while $\succeq$ ranks options.

We have just explained how to simplify the semantics $REAS$: by its equivalence to $REAS + Univ^*$, we may restrict attention to full-relevance universalist interpretations, in which N drops out as a parameter and the reasons structure reduces to a relation $\geq$ over property sets. An analogous simplification move works for $REAS + Cons + Univ$: by its equivalence to $REAS + Cons^*$, we may restrict attention to full-relevance consequentialist interpretations, in which again N drops out as a parameter. But this time the underlying assumption is that *all and only option* properties matter, so that the reasons structure boils down to a relation $\geq$ over *option* property sets.

A similar simplification move fails to exist for many other variants of reason-based semantics. For instance, in $REAS + Cons$ and $REAS + Moni$ we cannot restrict attention to a fixed and constant normative relevance function N. That is, the possibility of relativism here genuinely adds something to the semantics in question, by weakening it.

The deontic equivalence of every moral theory (reasons structure) to a universalist one by no means implies that all moral theories are 'essentially universalist'. Relativism is not made redundant as a potential meta-ethical position. The universalist counterpart of a relativistic theory (reasons structure) constitutes a genuinely different moral theory, although it happens to deliver the same deontic verdicts. It offers an entirely different (and possibly

[15] This ranking is defined as follows: for all $x, y \in X$, $x \succeq y \Leftrightarrow N(x,K) \geq N(y,K)$, where the choice of context K is arbitrary since $N(x,K)$ and $N(y,K)$ do not depend on K by consequentialism and universalism.

artificial and unconvincing) account of justification, reasons, or right-making features. Correctly interpreted, Remark 8 – far from making relativism irrelevant – highlights an underdetermination problem: the underdetermination of moral theory by deontic content, which has already been stressed by different authors (e.g., Broome 2004, Parfit 2011, Dietrich and List 2017).

6 Appendix: proofs

6.1 Proof of Theorem 4

Lemma 11. *Any behavioural interpretation is equivalent to a basic interpretation satisfying Ai, and vice versa.*

Lemma 12. *A basic interpretation satisfying Ai validates $Ax1$ at each world, and a basic interpretation validating $Ax1$ at each world is equivalent to a basic interpretation satisfying Ai.*

Proof of Lemma 11. First, any behavioural interpretation (W, v, R) is equivalent to the basic interpretation (W, v, ρ) in which $w \rho w'$ if and only if $K_w = K_{w'}$ and $x_{w'} \in R(K_w)$ $(= R(K_{w'}))$. Note that ρ satisfies Ai. Conversely, if a basic interpretation (W, v, ρ) satisfies (Ai), then we can define a rightness function by assigning to each context K the set $R(K) = \{x_{w'} : w' \in \rho(w)\}$, where w is some (hence by Ai any) world with context K. This guarantees that $R(K) \subseteq [K]$, so that R is a well-defined rightness function. One easily checks equivalence of the behavioural interpretation (W, v, R) to (W, v, ρ). ■

The proof of Lemma 12 draws on a simple lemma:

Lemma 13. *For each sentence ψ, $\vDash_{BASIC} \Box\psi \to \psi$. For each sentence ϕ it follows (via $\psi = \mathbf{O}\phi$) that$\vDash_{BASIC+Ax1} \Box\mathbf{O}\phi \leftrightarrow \mathbf{O}\phi$.*

Proof. Left to the reader.

Proof of Lemma 12. First, consider a basic interpretation $\mathcal{M} = (W, v, \rho)$ satisfying Ai, and a sentence $\phi \in \mathcal{L}$. To show that $\mathbf{O}\phi \Rightarrow \Box\mathbf{O}\phi$ holds at each world, we consider a $w \in W$ such that $\mathcal{M}, w \vDash \mathbf{O}\phi$ and have to show that $\mathcal{M}, w \vDash \Box\mathbf{O}\phi$, i.e., that $\mathcal{M}, w' \vDash \mathbf{O}\phi$ for all worlds w' with same context as w. Let w' be such a world. So $\rho(w) = \rho(w')$ by Ai. Hence, as $\mathcal{M}, w \vDash \mathbf{O}\phi$, we have $\mathcal{M}, w' \vDash \mathbf{O}\phi$.

Conversely, now consider a basic interpretation $\mathcal{M} = (W, v, \rho)$ validating $Ax1$ at each world. For each context $K \in \mathcal{K}$ fix an option $x_K \in [K]$ (strictly speaking, the existence of such an assignment $K \mapsto x_K$ assumes the axiom of choice). Define a new basic interpretation $\mathcal{M}' = (W, v, \rho')$ with same W and v as follows: for each world $w = (x, K)$ let $\rho'(x, K) = \rho(x_K, K)$. By construction, $\mathcal{M}'$ satisfies Ai. We now show that $\mathcal{M}$ is equivalent to $\mathcal{M}'$. Concretely, we have to show the following for each $\phi \in \mathcal{L}$:

$$\mathcal{M}, w \vDash \phi \Leftrightarrow \mathcal{M}', w \vDash \phi \text{ for all worlds } w \in W. \tag{1}$$

We prove (1) by induction on ϕ. First, if ϕ is atomic, i.e., in $\mathcal{P}$ or X, then (1) holds because $\mathcal{M}$ and $\mathcal{M}'$ share the truth function v on $W \times \mathcal{P}$. Now assume (1) holds for ϕ and ψ. Then:

- (1) holds for $\neg\phi$ since for all worlds w we have $\mathcal{M}, w \nvDash \phi \Leftrightarrow \mathcal{M}', w \nvDash \phi$, whence $\mathcal{M}, w \vDash \neg\phi \Leftrightarrow \mathcal{M}', w \vDash \neg\phi$.

- (1) holds for $\phi \wedge \psi$ since at all worlds w we have $\mathcal{M}, w \vDash \phi, \psi \Leftrightarrow \mathcal{M}', w \vDash \phi, \psi$, so that $\mathcal{M}, w \vDash \phi \wedge \psi \Leftrightarrow \mathcal{M}', w \vDash \phi \wedge \psi$.

- Next consider $\Box\phi$, and any world w. By assumption $\mathcal{M}, \bar{w} \vDash \phi \Leftrightarrow \mathcal{M}', \bar{w} \vDash \phi$ for all worlds $\bar{w}$. So $[\mathcal{M}, \bar{w} \vDash \phi$ for all $\bar{w}$ with $K_{\bar{w}} = K_w] \Leftrightarrow [\mathcal{M}', \bar{w} \vDash \phi$ for all $\bar{w}$ with $K_{\bar{w}} = K_w]$, i.e., $\mathcal{M}, w \vDash \Box\phi \Leftrightarrow \mathcal{M}', w \vDash \Box\phi$.

- Now consider $\phi \Rightarrow \psi$ and a world w. Since, for all worlds $\bar{w}$, $\mathcal{M}, \bar{w} \vDash \phi \Leftrightarrow \mathcal{M}', \bar{w} \vDash \phi$ and $\mathcal{M}, \bar{w} \vDash \psi \Leftrightarrow \mathcal{M}', \bar{w} \vDash \psi$, we have that [for all worlds $\bar{w}$ if $\mathcal{M}, \bar{w} \vDash \phi$ then $\mathcal{M}, \bar{w} \vDash \psi$] $\Leftrightarrow$[for all worlds $\bar{w}$ if $\mathcal{M}', \bar{w} \vDash \phi$ then $\mathcal{M}', \bar{w} \vDash \psi$]. In other words, $\mathcal{M}, w \vDash \phi \Rightarrow \psi \Leftrightarrow \mathcal{M}', w \vDash \phi \Rightarrow \psi$.

- Finally, and less trivially, consider $\mathbf{O}\phi$ and a world $w = (x, K)$. We must show that $\mathcal{M}, w \vDash \mathbf{O}\phi \Leftrightarrow \mathcal{M}', w \vDash \mathbf{O}\phi$. Note that

$$\begin{aligned} \mathcal{M}, w \vDash \mathbf{O}\phi &\Leftrightarrow \mathcal{M}, w \vDash \Box\mathbf{O}\phi && \text{by Lemma 13} \\ &\Leftrightarrow \mathcal{M}, w_K \vDash \Box\mathbf{O}\phi && \text{as } K_w = K_{w_K} \ (= K) \\ &\Leftrightarrow \mathcal{M}, w_K \vDash \mathbf{O}\phi && \text{by Lemma 13.} \end{aligned}$$

So it remains to show that $\mathcal{M}, w_K \vDash \mathbf{O}\phi \Leftrightarrow \mathcal{M}', w \vDash \mathbf{O}\phi$, i.e., that $[\mathcal{M}, \bar{w} \vDash \phi$ for all $\bar{w} \in \rho(w_K)]\Leftrightarrow [\mathcal{M}', \bar{w} \vDash \phi$ for all $\bar{w} \in \rho'(w)]$. The latter holds because $\rho(w_K) = \rho'(w)$ and because by induction hypothesis $\mathcal{M}, \bar{w} \vDash \phi \Leftrightarrow \mathcal{M}', \bar{w} \vDash \phi$ for all worlds $\bar{w}$.■

6.2 Proof of Theorem 5

Lemma 14. *Any ranking-based interpretation is equivalent to a behavioural interpretation satisfying Rev, and vice versa.*

Lemma 15. *If X is finite, a behavioural interpretation satisfies Rev if and only if it validates $Ax2$ at each world.*

Proof of Lemma 14. Any ranking-based interpretation $(W, v, \succeq)$ is equivalent to the behavioural interpretation (W, v, R) in which R is the rightness function induced by $\succeq$. This R satisfies *Rev* by Richter's Theorem (in the generalized version of Dietrich-List 2016, 2017). Conversely, if a behavioural interpretation (W, v, R) satisfies *Rev*, then, again by that theorem R is induced by some binary relation $\succeq$ on X, and (W, v, R) is clearly equivalent to $(W, v, \succeq)$. ■

The proof of Lemma 15 begins with a simple observation:

Lemma 16. *For all $x, y \in X$ and behavioural interpretations $\mathcal{M} = (W, v, R)$, x sometimes beats y w.r.t. R if and only if $SB(x, y)$ is true at all worlds of $\mathcal{M}$.*

Proof. Consider $x, y \in X$ and a behavioural interpretations $\mathcal{M} = (W, v, R)$. First assume $\mathcal{M}, w' \vDash SB(x, y)$ for all $w' \in W$. Then for some world w we have $\mathcal{M}, w \vDash \mathbf{P}x$ and $\mathcal{M}, w \vDash \diamond y$, or equivalently $x \in R(K_w)$ and $y \in [K_w]$. So x sometimes beats y w.r.t. R. Conversely, assume x sometimes beats y. Then we may pick a context K such that $x \in R(K)$ and $y \in [K]$. So$\mathcal{M}, w \vDash \mathbf{P}x$ and $\mathcal{M}, w \vDash \diamond y$ for worlds w with context K. So $\mathcal{M}, w' \vDash SB(x, y)$ for all $w' \in W$. ■

Proof of Lemma 15. Let X be finite, and consider a behavioural interpretation $\mathcal{M} = (W, v, R)$.

First assume R satisfies *Rev*, and consider a choice sentence $x \in X$. To show that $(\diamond x \wedge [\wedge_{y \in X}(\diamond y \rightarrow SB(x, y))]) \Rightarrow \mathbf{P}x$ holds at all worlds of $\mathcal{M}$, consider a world $w = (z, K)$ such that $\mathcal{M}, w \vDash \diamond x \wedge [\wedge_{y \in X}(\diamond y \rightarrow SB(x, y))]$, and let us show that $\mathcal{M}, w \vDash \mathbf{P}x$. By assumption, $\mathcal{M}, w \vDash \diamond x$, and $\mathcal{M}, w \vDash \diamond y \rightarrow SB(x, y)$ for all $y \in X$. So $x \in [K]$, and for all $y \in X$, if $y \in [K]$, then $\mathcal{M}, w \vDash SB(x, y)$, meaning by Lemma 16 that x sometimes beats y. Since each $y \in [K]$ is sometimes beaten by x, we have $x \in R(K)$ by *Rev*. So, $\mathcal{M}, w \vDash \mathbf{P}x$.

Conversely, suppose $\mathcal{M}$ validates $Ax2$ at each world. To show Rev, consider a context $K \in \mathcal{K}$ and a feasible option $x \in [K]$ such that each $y \in [K]$ is sometimes beaten by x. We have to show that $x \in R(K)$. Pick a world w with context K. By assumption, $\mathcal{M}, w \vDash \diamond x$ (as $x \in [K]$), and, for each $y \in X$, if $\mathcal{M}, w \vDash \diamond y$ (i.e., if $y \in [K]$), then $\mathcal{M}, w \vDash SB(x, y)$ (as this means that x sometimes beats y by Lemma 16). So $\mathcal{M}, w \vDash (\diamond x \wedge [\wedge_{y \in X}(\diamond y \to SB(x, y))])$. Since $\mathcal{M}$ validates $Ax2$ at w, it follows that $\mathcal{M}, w \vDash \mathbf{P}x$. So $x \in R(K)$. ■

6.3 Proof of Theorem 6

Lemma 17. *Any reason-based interpretation is equivalent to a behavioural interpretation satisfying $Re1$ and $Re2$, and vice versa.*

Lemma 18. *If $\mathcal{P}$ is finite, a behavioural interpretation satisfies $Re1$ if and only if it validates $Ax3$ at each world.*

Lemma 19. *If $\mathcal{P}$ is finite, a behavioural interpretation satisfies $Re2$ if and only if it validates $Ax4$ at each world.*

Proof of Lemma 17. Each reason-based interpretation $(W, v, \mathcal{R})$ is equivalent to the behavioural interpretation (W, v, R) in which R is the rightness function induced by $\mathcal{R}$. This R satisfies $Re1$ and $Re2$ by Dietrich-List (2016, 2017). Conversely, if a behavioural interpretation (W, v, R) satisfies $Re1$ and $Re2$, then by the same theorem $\mathcal{R}$ is induced by some reasons structure $\mathcal{R}$, and (W, v, R) is obviously equivalent to $(W, v, \mathcal{R})$. ■

Notation: If $\mathcal{P}$ is finite, then to each set $S \subseteq \mathcal{P}$ corresponds a sentence in $\widehat{\mathcal{P}}$ which is defined and denoted as $p_S = (\wedge_{s \in S} s) \wedge (\wedge_{s \in \mathcal{P} \setminus S} \neg s)$; so $\widehat{\mathcal{P}} = \{p_S : S \subseteq \mathcal{P}\}$.

Lemma 20. *Let $\mathcal{P}$ be finite. For all behavioural interpretations $\mathcal{M} = (W, v, R)$, sets $S \subseteq \mathcal{P}$, and contexts K,*

- *S is feasible in context K if and only if $\mathcal{M}, w \vDash \diamond p_S$ for worlds w with context K,*
- *S is right in context K (w.r.t. R) if and only if $\mathcal{M}, w \vDash \mathbf{P}p_S$ for worlds w with context K.*

Proof. Let $\mathcal{P}$, $\mathcal{M} = (W, v, R)$, S and K be as specified. First, S is feasible in K if and only if there is an $x \in [K]$ such that $S = \mathcal{P}(x, K)$, i.e., such that $\mathcal{M}, (x, K) \vDash p_S$. Equivalently, $\mathcal{M}, w \vDash \diamond p_S$ for worlds with context K. Second, S is right in K if and only if there is an $x \in R(K)$ such that $S = \mathcal{P}(x, K)$, i.e., such that $\mathcal{M}, (x, K) \vDash p_S$. Equivalently, $\mathcal{M}, w \vDash \mathbf{P}p_S$ for worlds with context K. ■

Proof of Lemma 18. Let $\mathcal{P}$ be finite, and $\mathcal{M} = (W, v, R)$ a behavioural interpretation.

First assume R satisfies $Re1$, and $S \subseteq \mathcal{P}$. To show that $(\mathbf{P}p_S \wedge \diamond(p_S \wedge x)) \Rightarrow \mathbf{P}x$ holds at all worlds of $\mathcal{M}$, consider any world w and assume $\mathcal{M}, w \vDash \mathbf{P}p_S \wedge \diamond(p_S \wedge x)$. We show that $\mathcal{M}, w \vDash \mathbf{P}x$. Let K be w's context. As $\mathcal{M}, w \vDash \mathbf{P}p_S$, S is right in context K by Lemma 20. So $S = \mathcal{P}(y, K)$ for some right option $y \in R(K)$. Meanwhile, as $\mathcal{M}, w \vDash \diamond(p_S \wedge x)$ there is a $z \in [K]$ such that $\mathcal{M}, (z, K) \vDash p_S \wedge x$. It follows that $z = x$, and so that $\mathcal{M}, (x, K) \vDash p_S$. Hence, $\mathcal{P}(x, K) = S$. Since $\mathcal{P}(x, K) = \mathcal{P}(y, K)$ and since $y \in R(K)$, we have $x \in R(K)$ by $Ax3$.

Conversely, assume $\mathcal{M}$ validates $Ax3$ at all worlds. To show $Re1$, consider a context K, a feasible option $x \in [K]$, and a right option $y \in R(K)$ such that $\mathcal{P}(x, K) = \mathcal{P}(y, K)$. We must show that $x \in R(K)$. Write S for the set $\mathcal{P}(x, K) = \mathcal{P}(y, K)$. Let w be any world with context K. As $S = \mathcal{P}(y, K)$ and $y \in R(K)$, the set S is right in context K. So (*) $\mathcal{M}, w \vDash \mathbf{P}p_S$ by Lemma 20. Further, $\mathcal{M}, (x, K) \vDash p_S$ (as $\mathcal{P}(x, K) = S$) and $\mathcal{M}, (x, K) \vDash x$. So $\mathcal{M}, (x, K) \vDash p_S \wedge x$, and hence (**) $\mathcal{M}, w \vDash \diamond(p_S \wedge x)$. By (*) and (**) and the validity of $Ax3$ at all worlds, we have $\mathcal{M}, w \vDash \mathbf{P}x$. Hence $x \in R(K)$. ■

Finally, to prove Lemma 19 we first establish a simple fact (analogous to Lemma 16):

Lemma 21. *Let $\mathcal{P}$ be finite. For all $S, S' \subseteq \mathcal{P}$ and behavioural interpretations $\mathcal{M} = (W, v, R)$, S sometimes beats S' w.r.t. R if and only if $SB(p_S, p_{S'})$ is true at all worlds of $\mathcal{M}$.*

Proof. Let $\mathcal{P}$, S, S', and $\mathcal{M} = (W, v, R)$ be as specified. Now $SB(p_S, p_{S'})$ holds at all worlds of $\mathcal{M}$ if and only if $\mathcal{M}, w \vDash \mathbf{P}p_S$ and $\mathcal{M}, w \vDash \diamond p_{S'}$ for some world w. Equivalently (by Lemma 20) S is right and S' is permissible in some context, i.e., S sometimes beats S'. ■

Proof of Lemma 19. Let $\mathcal{P}$ be finite, and $\mathcal{M} = (W, v, R)$ be a behavioural interpretation.

Suppose first R satisfies $Re2$, and $S \subseteq \mathcal{P}$. We must show that $(\diamond p_S \wedge [\wedge_{S' \subseteq \mathcal{P}}(\diamond p_{S'} \rightarrow SB(p_S, p_{S'}))]) \Rightarrow \mathbf{P}p_S$ holds at all worlds of $\mathcal{M}$. Let $w =$

(z, K) such that $\mathcal{M}, w \vDash \diamond p_S \wedge [\wedge_{S' \subseteq \mathcal{P}} (\diamond p_{S'} \rightarrow SB(p_S, p_{S'}))]$. We show that $\mathcal{M}, w \vDash \mathbf{P} p_S$. By assumption, $\mathcal{M}, w \vDash \diamond p_S$, and, for each $S \subseteq \mathcal{P}$, if $\mathcal{M}, w \vDash \diamond p_{S'}$ then $\mathcal{M}, w \vDash SB(p_S, p_{S'})$. By Lemmas 20 and 21 this means that S is feasible in context K and, for each $S' \subseteq \mathcal{P}$, if S' is feasible in K then S sometimes beats S'. Hence S is right in K by $Re2$, so that $\mathcal{M}, w \vDash \mathbf{P} p_S$ by Lemma 20.

Conversely, suppose $\mathcal{M}$ validates $Ax4$ at each world. To show $Re2$, assume that $S \subseteq \mathcal{P}$ is feasible in a given context K and each $S' \subseteq \mathcal{P}$ feasible in K is sometimes beaten by S. By Lemmas 20 and 21 this means that, at worlds w with context K, $\mathcal{M}, w \vDash \diamond p_S$ and, for all $S \subseteq \mathcal{P}$, if $\mathcal{M}, w \vDash \diamond p_{S'}$ then $\mathcal{M}, w \vDash SB(p_S, p_{S'})$. So, still at worlds w with context K, $\mathcal{M}, w \vDash (\diamond x \wedge [\wedge_{y \in X} (\diamond y \rightarrow SB(x, y))])$. Since $\mathcal{M}$ validates $Ax2$ at all worlds, it follows that $\mathcal{M}, w \vDash \mathbf{P} x$. So $x \in R(K)$ by Lemma 20. ■

References

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

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

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

Dietrich, F., List, C. (2013a) A reason-based theory of rational choice, *Noûs* 47(1): 104-134

Dietrich, F., List, C. (2013b) Where do preferences come from? *International Journal of Game Theory* 42(3): 613-637

Dietrich, F., List, C. (2016) Reason-based choice and context-dependence: an explanatory framework, *Economics and Philosophy* 32(2): 175-229

Dietrich, F., List, C. (2017) What matters and how it matters: a choice-theoretic representation of moral theories, *Philosophical Review*, forthcoming

Gabbay, D., Horty, J., Parent, X. et al. (eds.) (2013) *Handbook of Deontic Logic and Normative Systems*, London: College Publications

Hansson, S. O. (2001) Preference logic. In D. Gabbay and F. Guenthner, eds., *Handbook of Philosophical Logic*, vol. 4, pp. 319-393. Dordrecht: Kluwer

Liu, F. (2010) Von Wright's 'The Logic of Preference' revisited, *Synthese 175*: 69-88

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

Priest, G. (2001) An Introduction to Non-classical Logic, Cambridge Univ. Press

Richter, M. (1971) Rational Choice. In: *Preferences, Utility, and Demand*, ed. J. S. Chipman et al., 29–58. New York: Harcourt Brace Jovanovich
Samuelson, P. (1948) Consumption theory in terms of revealed preferences, *Economica* 15: 243–253
Segerberg, K., Meyer, J.-J., Kracht, M. (2013) The Logic of Action. In: *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), E. N. Zalta (ed.). URL: https://plato.stanford.edu/archives/win2016/entries/logic-action
Sen, A. K. (1993) Internal Consistency of Choice, *Econometrica* 61: 495–521