Contents

Research Articles

Organon F 27 (2) 2020: 142–168
https://doi.org/10.31577/orgf.2020.27201

ISSN 2585-7150 (online)
ISSN 1335-0668 (print)

RESEARCH ARTICLE

Unification and the Myth of Purely Reductive Understanding

Michael J. Shaffer*

Received: 20 April 2019 / Accepted: 6 August 2019

ABSTRACT: In this paper significant challenges are raised with respect to the view that explanation essentially involves unification. These objections are raised specifically with respect to the well-known versions of unificationism developed and defended by Michael Friedman and Philip Kitcher. The objections involve the explanatory regress argument and the concepts of reduction and scientific understanding. Essentially, the contention made here is that these versions of unificationism wrongly assume that reduction secures understanding.

KEYWORDS: Explanation; reduction; simplicity; scientific understanding; unification.

1. Introduction

In post-Hempelian discussions of the nature of scientific explanation it is commonplace to note that one of the main functions of such explanation is to yield understanding, more specifically, scientific understanding.[1] That

[1] See (de Regt 2017) and (de Regt, et al. 2009).

* St. Cloud State University
✐ Department of Philosophy, St. Cloud State University, CH365N, 720 4th Ave. S., St. Cloud, MN 56301, USA
✉ mjshaffer@stcloudstate.edu

Hempel's covering-law model(s) of explanation failed to adequately account for this desideratum is one of the reasons that motivated the general rejection of the Hempelian account(s) of explanation.[2] It is apparent then that a theory of scientific explanation that does not show how explanations of these sorts yield such understanding is *ipso facto* inadequate. We can refer to this constraint on theories of explanation as *the understanding thesis.*

This paper explores how this constraint on one of the most prominent theories of explanation is supposed to be satisfied. In particular, it focuses on one of the most widely held successor theories to Hempel's deductive-nomological and inductive-statistical models of explanation: the view that explanation is a sort of global unification.[3] This is the view that explanation is essentially achieved when a phenomenon or law is integrated into the simplest global system that organizes (or systematizes) our corpus of beliefs. In other words, explanation is a matter of showing how various, apparently unrelated, beliefs can be derived logically from a small set of premises or axioms, thereby reducing the number of beliefs that must be accepted as brute axiomatic beliefs. The argument presented here concludes that standard unificationist theories of explanation of this sort, like the most well-known versions proposed by Michael Friedman (1974) and Philip Kitcher (1981, 1989, 1993), fail to convincingly show how unification yields understanding in any interesting, non-trivial, sense.

Establishing this result involves seriously contending with a particular traditional objection to explanation that has often been blithely ignored. This is *the explanatory regress argument* (ERA). This wide-scoped objection to accounts of explanation is based on the observation that explanations

[2] See (Salmon 1989) for the canonical recounting of the relevant history of the concept of explanation, and (Lambert 1980) for some additional criticisms of the view that Hempel's covering-law models yield understanding. (Woodward 2017) is also a useful survey of theories of explanation.

[3] Kitcher (1981) points out that lurking behind the official, logical empiricist, doctrine espoused by Hempel was a version of the unificationist view he advocates. This implicit view can be found in (Hempel 1965, 1966) and in Feigl's classic (1970). Prominent contemporary defenders of this view include Friedman (1974) and Kitcher (1981, 1989, 1993).

involve relating the unfamiliar to the familiar, or relating that which is not understood to that which is.[4] According to the ERA bona fide explanation occurs only when that which is being explained—the explanandum—is explained by something—the explanans—that is already itself explained.[5] This point can then be related to scientific understanding by noting that understanding is, in an important sense, parasitic on explanation. Understanding can be achieved only by relating what we do not understand to what we do understand. But then (so the criticism goes) we are faced with what appears to be a regress that can be terminated only by the positing of some privileged body of explanatorily primitive unexplained explainers that themselves generate all understanding. As there do not appear to be such fundamental privileged explainers, it seems that if anything is explained at all and anything understood, explanation *must not* require that an explanandum can only be explained by an explanans that is itself already explained. Unificationists often portray themselves as being able to avoid this problem. For example, Friedman explicitly tells us that,

> This global view of scientific understanding [unificationism] also, it seems to me, provides the correct answer to the old argument that science is incapable of explaining anything because the basic phenomena to which others are reduced are themselves neither explained nor understood. According to this argument, science merely transfers our puzzlement from one phenomenon to another; it replaces one surprising phenomenon by another equally surprising phenomenon. (Friedman 1974, 18)

However, here it will be shown that unificationist views like Friedman's do not provide an adequate answer to the ERA, and do not provide an acceptable account of the relation between explanation and understanding.

[4] This is especially true of (Friedman 1974).

[5] The real worry behind the ERA appears to be that if we do not require that the explanans is itself explained, we will be forced to posit unexplained explainers to terminate the regress or we must argue that explanation is somehow generated *sui generis* by the explanation relation. In ignoring the ERA, and adopting a global view of explanation the unificationists seem to favor the second option.

It will be argued that in attempting to avoid the ERA and replace Hempel's covering-law model of explanation, unificationist views fail to connect explanation to understanding in any interesting, non-trivial, sense. The failure on the part of the unificationists to adequately link explanation to understanding non-trivially, is in part the result of the unrealistic epistemic nature of the unificationist view and the acceptance of two dogmatic views about the relationships between simplicity, truth and understanding. To put the point generally, the unificationist view of explanation requires epistemic agents and communities to possess computational resources that far exceed those that are available, and so cannot possibly yield a realistic account of scientific understanding or it leads to skepticism about explanation. Since neither of these views is acceptable, it will be suggested that the unificationist view should be rejected.

On a more positive note, it will be suggested that examination both of the ERA and of the criticisms of the unificationist view reveals that there are two distinct but equally important concepts of scientific understanding that should be distinguished. The first concept is that of *semantic understanding*. The second is that of *reduction*. When these different senses of 'understanding' are properly distinguished, it is possible to foresee the construction of a complete, non-skeptical and naturalistically acceptable account of explanation that can truly yield an acceptable account of scientific understanding.

2. Explanation as unification

Before turning to the critical appraisal of the unificationist view, we must look at the details of the unificationist theory of explanation. As already noted, the two most prominent versions of unificationism are those independently offered by Michael Friedman and by Philip Kitcher. So, first, we need to examine the basic details of these views with an eye to determining how Kitcher's and Friedman's specific unificationist views of explanation are supposed to yield scientific understanding, as per the understanding thesis. Note, however, that sections 2.3 and 2.4 contain presentations of the technical details, respectively, of Friedman's and Kitcher's views of explanatory unification and these sections can be skipped or skimmed over for those who are not especially interested in the technicalities involved. Ultimately, as we shall see, what is most important is that we note that both Friedman and Kitcher subscribe to a core set of views about explanation that involves unification and simplification that is supposed to satisfy the understanding thesis. Let us then begin by looking at the informal versions of Friedman's and Kitcher's views of explanatory unification.

2.1. Friedman, Kitcher and unificationism

Friedman presents a version of the unificationist view that is based on the idea that what we are inclined to regard as being in need of explanation are laws and not ordinarily events, *pace* Hempel. More crucially, explanation of laws is achieved by showing that the phenomenon described by the target law are really just cases of some other phenomenon described by a more fundamental law. This is just the familiar relation of reduction, and Friedman candidly tells us that,

> The central problem for the theory of explanation comes down to this: what is the relation between phenomena in virtue of which one phenomenon can constitute an explanation of another, and what is it about this relation that gives understanding of the explained phenomenon. (Friedman 1974, 6)

So for unificationists like Friedman it would seem to be the case that explanation, and thereby understanding, is supposed to be achieved by reduction. Friedman motivates the consideration of unificationism as a serious option by rejecting as inadequate Hempel's covering-law model, and what he, respectively, calls the familiarity and intellectual fashions views of explanation.

The intellectual fashions view will be ignored here as it is irrelevant to the issue raised in this paper, but the familiarity view will play a more important role in the arguments to follow. Essentially the familiarity view holds that explanation is a matter of relating the unexplained to the explained by relating the unfamiliar to the familiar. Friedman's specific criticisms of these views will not be repeated here. However, in criticizing these two views, Friedman importantly establishes a set of three desiderata that any adequate theory of explanation should satisfy:

(DE1) The theory should be sufficiently general.

(DE2) The theory should be objective.

(DE3) The theory should connect explanation and understanding (Friedman 1974, 13-14).[6]

It is clear from the passage quoted in section 1 that Friedman is especially concerned to show that his own account of explanation is capable of satisfying DE3, but he is also careful to explain that,

> When I ask that a theory of scientific explanation tell us what it is about the explanation relation that produces understanding, I do not suppose that 'scientific understanding' is a clear notion. Nor do I suppose that it is possible to say what scientific understanding is in advance of giving a theory of explanation. It is not reasonable to require that a theory of explanation proceed by first defining 'scientific understanding' and then showing how its reconstruction of the explanation relation produces scientific understanding. We can find out what scientific understanding consists in only by finding out what scientific explanation is and vice versa. (Friedman 1974, 6)

This disclaimer is rather troubling, and this passage contains an anticipation of the criticism that unificationism only trivially satisfies DE3. Later we will return to this issue, but, at this point, it suffices to note that Friedman sees traditional views of explanation as all failing to satisfy one or more of DE1-DE3 and argues that unificationism does not suffer from these problems.

In the place of the more traditional theories of explanation, Friedman then offers us unificationism as a replacement. Friedman's unificationist theory depends on the following core intuition:

> Science increases our understanding of the world by reducing the total number of independent phenomena that we have to accept as ultimate or given. A world with fewer phenomena is, other things equal, more comprehensible than one with more. (Friedman 1974, 15)

[6] DE3 is, of course, just the understanding thesis mentioned in the introduction.

Friedman's views about explanation then essentially include the following ideas. First, explanation is unification. Second, the best unification is the simplest one. Finally, explanation yields understanding because the simplest unification is the most understandable one. This last point is crucial to note. On Friedman's view the best unification is understandable because it requires us to accept the total systematization of our beliefs about some body of phenomena that involves the fewest primitive beliefs.

Having laid out the details of Friedman's view we can now turn our attention to Kitcher's view, and it is interesting to note that Kitcher's view was constructed as a sympathetic, but critical, reaction to the deficiencies of Friedman's unificationist analysis of explanation.[7] Kitcher's unificationism is then importantly similar to that of Friedman. Most crucially, it is clear that Kitcher and Friedman have (at least roughly) the same sort of idea in mind concerning the issue of how unification yields explanation, and thereby, understanding. Specifically, it is by the reduction of our set of our antecedently accepted beliefs, K, to the *simplest* systematization of those beliefs that is supposed to issue in greater understanding. This is ultimately due to the simplicity of the reducing set of beliefs. In his (1989) Kitcher explicitly introduces the following principle of acceptance for systematic unifications:

(U) S should be chosen over S' as the explanatory store over K, $E(K)$, just in case S has greater unifying power with respect to K than S' (Kitcher 1989, 477).

Essentially what Kitcher advocates is that we accept the global theory that *best* unifies, that best explains, the totality of our antecedent beliefs by showing that they can be derived from a small set of special argument patterns involving a small set of basic (axiomatically accepted) beliefs that, in part, constitute the reducing theory.

As with Friedman's view what we have here is a form of inference to the best explanation (IBE), where the quality, or 'bestness', of an explanation is to be measured in terms of its ability to simplify our global belief system and thus supposedly generate understanding. Kitcher explicitly tells us,

[7] In fact, (Kitcher 1976) is intended to be an explicit criticism of Friedman's account of explanatory unification.

> I have sketched an account of explanation as unification, attempting to show that such an account has the resources to provide insight into episodes in the history of science and to overcome some traditional problems for the covering law model. In doing so, let me indicate very briefly how my view of explanation as unification suggests how scientific explanation yields understanding. By using a few patterns of argument in the derivation of many beliefs we minimize the number of *types* of premises we must take as underived. That is, we reduce, in a far as is possible, the number of types of facts we must accept as brute. Hence we can endorse something close to Friedman's view of the merits of scientific explanation. (Kitcher 1981, 529)

So, as with Fiedman's unificationism, such a reduction is supposed to make our global belief system more understandable by reducing the number of independent facts we must accept as basic, and so Kitcher also endorses DE3.

2.2. Friedman's formal account of unification[8]

For those interested in the technical details, Friedman gives this informal idea a formal treatment as follows. We are to regard a scientific community C, at a given time t, as accepting a set of law-like sentences K. In other words, K is the set of laws accepted in C at t. K is to be understood to be deductively closed such that if S is a law-like sentence and $K \vdash S$, then $S \in K$. We are then supposed to accept the systematization of K that reduces K to the smallest set of independently acceptable sentences, where independent acceptability is characterized as follows:

i. If $S \vdash Q$, then S is not acceptable independent of Q.

ii. If S is acceptable independently of P and $Q \vdash P$, then S is acceptable independently of Q. (Friedman 1974, 16)

Reduction is to be understood formally in the following manner. We first define a *partition* of the sentence S as the set of sentences Γ such that Γ is

[8] As noted previously, readers who are not interested in the formal details of Friedman's view can skip this section.

logically equivalent to S and each S', where $S' \in \Gamma$, is acceptable independently of S. Moreover, a sentence will be *K-atomic* if it has no partition in the sense just defined, and a *K-partition* of a set of sentences Δ will be a set Γ of K-atomic sentences logically equivalent to Δ. The *K-cardinality* of a set Δ, abbreviated K-card (Δ), is defined as inf {card (Γ): Γ is a K-partition of Δ}. We then get one of the key concepts of Friedman's unificationist view. S *reduces* the set Δ, if and only if, K-card ($\Delta \cap \{S\}$) > K-card (Δ). He then tells us that what we really want reduced is the number of independently acceptable consequences of S, $\mathrm{con}_K(\mathrm{S})$.

Friedman then ultimately defines *explanation* as follows:

(DI′) S_1 explains S_2 if and only if there exists a partition Γ of S_1 and an $S_i \in \Gamma$ such that $S_2 \in \mathrm{con}_K(S_i)$ and S_i reduces $\mathrm{con}_K(S_i)$. (Friedman 1974, 17)

On this basis, Friedman presents us with a relatively simple formal treatment of explanation as global unification that is supposed to yield understanding because global systematizations with fewer independent phenomena that we must understand primitively are more understandable. Presumably, we are supposed to accept the theory that best explains K for C at t, and the best explanation of K is that reduction of K in accord with DI´ that is globally the simplest reduction in the sense of simplicity just specified. Again, it is precisely because such unifications are simpler that they are supposed to be more understandable and this formal model shows how this works in terms of the concept of the simplest reduction.

2.3. Kitcher's alternative formal account of unificationism[9]

Kitcher too offers a formal account of the unificationist view of explanation, but it is presented in terms of a very different technical framework from that employed by Friedman. In any case, the main idea behind Kitcher's alternative version of unificationism is captured in the following passage:

[9] As noted previously, this section can be skipped by those who are not interested in the formal details of Kitcher's view.

> The general problem I have set is that of specifying $E(K)$, the *explanatory store over* K, which is the set of arguments acceptable as the basis for acts of explanation by those whose beliefs are exactly the members of K. (For purposes of this paper I shall assume that, for each K there is exactly on $E(K)$.
>
> The unofficial view [unificationism] answers the problem: for each K, $E(K)$ is the set of arguments which best unifies K. My task is to articulate this answer. (Kitcher 1981, 512)

It is important to note that Kitcher is focusing on the unification of K by appeal to a set of general argument patterns, and argument patterns are not argument forms in the sense employed in modern formal logic. As Kitcher understands them, argument patterns are more context-specific and less abstract than logical argument forms. Moreover, the reducing theory is not merely a set of sentences as in the case of Friedman's view. According to Kitcher each theory is constituted, in part, by a set of accepted argument styles, or canonical argument patterns that are regarded as 'good'.

Kitcher elaborates on this idea, and tells us that a *generating set* Σ is a set of argument patterns Π such that each argument in Σ instantiates a pattern in Π. A generating set for Σ is *complete with respect* to K, if and only if, every argument which is acceptable relative to K and which instantiates a pattern in Π is a member of Σ. To select the explanatory store $E(K)$ we first narrow our choice to those sets of arguments which are acceptable relative to K, the possible systematizations of K. From among the various generating sets of argument patterns that are complete with respect to K, in accordance with U we select that set with the greatest unifying power and we call the selected set the *basis* of the set of arguments in question. $E(K)$, the explanatory store over is K, that systematization whose basis does best by the criteria of unifying power (Kitcher 1981, 519-20). Importantly, this includes the idea that the best systematization is that which is simplest. So, let us then turn our attention to the central concept Kitcher's account explanation, the concept of unifying power.

Kitcher begins his explication of the concept of explanation as unification by offering a series of related definitions that are to be used in defining what he refers to as a general argument pattern. A *schematic sentence* is an expression generated by replacement of at least one of the non-logical expressions in the sentence with dummy letters. For each schematic sentence there must also be a set of *filling instructions* that specify how the dummy letters are to be replaced. A *schematic argument* is then a set of schematic sentences, and a *classification* for such an argument is a set of sentences describing the inferential features of the schematic argument. The classification, in effect, identifies which sentences are premises, which sentence is the conclusion, what logical rules of inference are used, etc. Kitcher then defines a *general argument pattern* as a sequence of sentences such that:

a. The sequence has the same number of terms as the schematic argument of the general argument pattern.

b. Each sentence in the sequence is obtained from the corresponding schematic sentence in accordance with the appropriate filling instructions.

c. It is possible to construct a chain of reasoning which assigns to each sentence the status accorded to the corresponding schematic sentence by the classification (Kitcher 1981, 516-17).

The task that Kitcher then undertakes is to explicate the concept of the unifying power of a set of argument patterns so that we can determine which set is the best explanatory unification of K.

After proposing that the unifying power of a set of argument patterns should be defined as the ability of those argument patterns to generate "…a large number of accepted sentences as the conclusions of acceptable arguments which instantiate a few, stringent patterns (Kitcher 1981, 520)," Kitcher points out that this suggestion will not work. Rather mysteriously, he then abandons the attempt to specify the concept of unifying power more precisely. Nonetheless, he does give us some idea of what he has in mind. First we define the conclusion set of a set of arguments Σ, as the set of sentences which occur as conclusions, $C(\Sigma)$, of some argument that is a member of Σ. Kitcher then suggests that the unifying power of a reduction base B_i with respect to K varies with the size of $C(\Sigma_i)$ and the stringency of the patterns of B_i, and inversely with the number of members of B_i (Kitcher 1981, 520). The idea is to select the smallest set of premises that generates the largest conclusion set with respect to our belief system.

The concept of stringency is itself left undefined, but Kitcher explains that the stringency of an argument pattern is determined by the strictness of the conditions governing the substitution of dummy letters in argument patterns and conditions governing acceptable logical structure (Kitcher 1981, 518). The former are fixed, for the most part, by the filling instructions, the latter by the classification. Considering the stringency of sets of argument patterns is introduced in order to rule out cases both where all argument patterns are acceptable relative to *K* and where only one unique argument is acceptable relative to *K*. In any case, Kitcher admits that this account of unifying is too simple and quite vague, but he offers nothing more well-developed as a replacement for this account of unifying power. Kitcher does note that, in addition to counting numbers of argument patterns, the concept of unifying power probably also ought to include in some way the similarities among such argument patterns, but he offers no deeper account of this concept.

2.4. Kitcher, Friedman and the common basis of unificationism

The unificationists we have considered (who are without any doubt the most well-known defenders of unificationism) hold that the best explanation, the one that we ought to accept, is the simplest systematization of our belief corpus. As it happens, this belief system is also supposed to be the one that affords us the most understanding. Kitcher and Friedamn are then clearly advocating a sophisticated form of IBE (or what Peirce variously called abduction or the method of hypothesis) on a global scale. More to the point, on their view the 'bestness' of an explanation is understood in terms of a kind of simplicity as it applies to reduction bases.[10] On this general point about the virtues of simplicity they are, in fact, in broad agreement with Peirce himself who often stressed the role of economy in abductive inference and in science in general, as is the case with most defenders of IBE.[11] But, they apply this notion of simplicity to complete belief systems as opposed to localized sets of beliefs. Consequently, the,

[10] We are here ignoring the differences between Peirce's conceptions of abduction and retroduction as developed in his later views.

[11] See, for example, Peirce (c. 1901/1931-1958, 6.529-6.5230 and 7.220).

the generic rule of theory acceptance such unificationists appear to endorse is:

(URA) Accept the best explanation of a total system of beliefs where the best explanation is the globally simplest unification of a global system of beliefs/sentences.

It is important to note, however, Peirce does not in any obvious way, seem to be committed to the idea that abduction is global in this sense, and this is the source of considerable trouble for unificationism.[12] In what follows, we will find that there is some reason to suspect that this kind of global IBE is not a computationally feasible form of inference. However, before turning our attention to the critique of unificationism, let us first look at one of its supposed successes.

2.5. The explanatory regress argument and unification

One of the virtues that unificationists have claimed for their theory is the avoidance of the ERA. As we saw in section 1, Friedman is especially clear about this virtue of his version of unificationism. His main objection to familiarity views like that defended by Dray (1964) and, to some degree, by Scriven (1970), is that we often explain laws by appeal to other laws with which we are less familiar. If this is so and we believe that these sorts of cases are *bona fide* instances of explanation, then explanation cannot be relating the unfamiliar to the familiar. Unification does not require relating the unfamiliar to the familiar and supposedly secures the connection between explanation and understanding by linking explanation with unification, and thereby simplicity. The unificationists avoid the ERA by ignoring it and simply jettisoning the requirement that we are more familiar with the explanans material than we are with the explanandum material.

Friedman is correct on this count. However, for the unificationists global unification in the simplest system then replaces any consideration of the kind of understanding discussed by familiarity theorists, but it is here where the unificationists go wrong. We *can* explain phenomena by relating them

[12] Some problems with treating inference to the best explanation globally are discussed in (Shaffer 2012) and in (Fodor 2000).

to unfamiliar theories. Simply consider most classical explanations of microscopic phenomena by quantum mechanics (QM). Friedman is correct that we are most assuredly *not* more familiar with the relevant aspects of the QM explanations. Nevertheless, we must possess an important kind of understanding of the reducing theory if such an explanation is to yield real understanding and so the familiarity theorists are correct to take the ERA seriously. In the particular case of QM, it is plausible to believe now that this principle is satisfied. *Pace* Friedman, it is not the *unfamiliarity* of reducing theories that is the real issue, but rather the issue is whether we *understand* the reducing theory in some relevant sense. On the one hand, all novel theories are unfamiliar, but his need not entail that they are not understood. On the other hand, familiar theories may not be such that they are understood. As such, it is a critical mistake to merely equate understanding with familiarity, and thus ignore the ERA as Friedman does.[13]

3. The failure of unificationism

So where does unificationism go wrong? As outlined in the introduction, unificationism fails for several closely connected reasons. First, in attempting to avoid the ERA unificationism places great, unwarranted, faith in the metaphysical view that our world is ultimately simple and in the view that the world can be successfully explained by a simple global theory. This is important because these particular articles of faith must be true for it to be possible for us to cognitively grasp entire, global, scientific belief systems (i.e. belief corpuses) that potentially apply to an ultimately comprehensible world. All of this must be the case because unificationists require that we be able to *compare* the simplicity of such explanatory systems and accept the simplest systematization of the observed facts about the world. But, it is not at all clear that our world is simple in this sense, and it is not computationally and epistemically feasible that we could (even as a community)

[13] The unificationists do not really avoid the ERA anyway, as they do not offer an account of how the basic beliefs are themselves understood.

compare global belief systems in the manner that unificationists require, even if global belief systems turn out to be relatively simple.

Supposing that we could manage to avoid these metaphysical and computational problems, it will be shown here that, *pace* the unificationists, the simplest unified belief corpus may not provide us with understanding in the relevant sense of the term and that the resulting unified system may not even be likely to be true. So, firstly, it will be argued here that unification is not sufficient for understanding and that the unificationists' account of explanation does not track the truth. In other words, given the unificationist view of explanation, it is perfectly possible that the best explanation (i.e. the simplest global unification of our scientific beliefs) is not understandable at all and may well be false. By DE3 (and DE2), it then follows that unificationism should be rejected if actual human understanding is factive. In any case, let us now turn our attention to the details of these criticisms of unificationism.

3.1. Two dogmas of simplicity

As the expositions of Friedman's and Kitcher's views show, the essential motivation behind unificationism appears to involve a particular notion of reduction that relates explanation to simplicity, and thereby explanation to understanding. This unificationist argument essentially involves three steps. The first step (S1) is to define explanation and best explanation in terms of the unification of belief systems. The second step (S2) is to note that such unification implies simplification. Finally, the third step (S3) is to argue that simpler theories are more understandable. The various elements of this line of reasoning are found in Kitcher's and Friedman's claims that we have examined in earlier sections and it is obvious that there is broad agreement between Friedman and Kitcher in this respect. This argument, however, depends on at least two crucial assumptions. First, the unificationists appear to accept the view that, other things equal, a numerically smaller system of beliefs and/or inference schemes is more understandable. Second, as Friedman and Kitcher both do not appear to be skeptics and seem to believe that we can successfully (i.e. truthfully) explain phenomena and thereby understand the world, it must be the case that they believe that the world really is, at least relatively, simple in some ontological sense.[14]

These assumptions are what we will call the *two dogmas of simplicity.* This first view, that a numerically smaller belief system is more understandable, is the *dogma of epistemic simplicity* (DES). The second dogma, the view that the world really is simple, as the *dogma of metaphysical simplicity* (DMS). As we shall see, both are mere dogmas and neither is non-controversially true. Nevertheless, their general acceptance–tacit or otherwise–has made unificationism rather more appealing than it otherwise would be, especially in light of their ignoring the ERA. DES is a crucial assumption behind S3 of the unificationist argument and DEM is, similarly, crucial for justifying S1 and S2.

3.2. The dogma of metaphysical simplicity

The unificationists' implied view that the world is simple appears either to be a matter of blind faith or it is an *a priori* matter, for surely it is not clearly the case on empirical grounds. However, as it is obvious that the fact of the matter concerning the simplicity of the world cannot be an *a priori* issue, the unificationists' assumption that the world is fundamentally simple appears merely to be a dogmatic assumption. In fact, empirical considerations seem to support the view that our world is rather complex, and that the world is divided into more than one independent domain characterized by entirely different, non-reducible, laws. Therefore, as neither appealing to faith nor to *a prioricity* is, at least in this case, legitimate, it seems that there is then no real substantive reason to suspect that our world really is simple. [15] Hence, we should have serious doubts that simple theories are more likely to be true, and this raises serious problems for unificationists.

Consider the following conditions of adequacy concerning explanation and IBE, all of which the unificationists appear to accept (implicitly or otherwise):

[14] If this were not the case, then it is unclear what the purpose of the many examples of allegedly successful explanation are.

[15] See (Fodor 1974; Dupré 1983; Cartwright 1999; and Shaffer 2012) for related points.

(CA1) The explanans of a *bona fide* explanation must bear some relation to the truth.[16]

(CA2) There is some connection between an explanation being the *best* explanation and that explanation's truth; the best explanation is likely to be true and,

(CA3) The 'bestness' of explanation is a function of its simplicity.[17]

If there is some connection between an explanation being the best explanation and that explanation's truth in this way (i.e. if CA2 and CA3 are correct in a way that allows CA1 to be satisfied), then, unless there is some non-dogmatic reason to believe DMS, the unificationists' view of the connection between unification and truth of explanation cannot be adequate. This is because they offer no reason to accept CA2 and CA3 except that doing so allows us to trivially link explanation and understanding in order to satisfy DA3. However, as it is not obviously the case that DMS is true and we do not have any justifications for accepting CA2 and CA3, we have no reason to suspect that CA1 is satisfied by the unificationists' views of explanation. As such, it is neither obviously true that simpler, more unified, global belief systems are explanatory, nor is it obviously true that simpler, more unified, global systems of belief are more likely to be true. So S1 and S2 of the unificationist argument are thus impugned.

3.3. The dogma of epistemic simplicity

In order to support S3 the unificationists argue that we ought to accept the simplest systematization of our scientific beliefs where the simplest systematization is the belief system that reduces the number of phenomena or inference patterns describing the phenomena to the smallest number by unification. As we have seen, the putative reason for doing so is that the simplest globally unified system of this sort is supposed to be the most understandable. Essentially, they believe that a system that is reductively simplified by unification is supposedly more understandable. This is an

[16] As is well known, Hempel and Oppenheim (1948) argue that the explanans must be true.

[17] Lipton (1991) and Thagard (1988) defend these conditions of adequacy for IBE.

article of pure epistemological faith (another dogma), and it is a view that is easily exploded if we take the ERA seriously, as the unificationists do not.

It seems obvious that a system that is unified in the unificationists' sense may be completely incomprehensible. Simply consider a belief system that is unified in this manner, but which reduces the apparent plethora of phenomena to a unified but relatively small (axiomatic) reduction base that is not itself understood at all. Should we really say that such a system provides us with an explanation that yields understanding? Surely not. The following well-understood sort of case demonstrates this and history is replete with examples of such non-explanatory reductive thinking. Suppose that a cognitive agent or a cognitive community attributes to all phenomena a relatively unified, but inscrutable, set of causes: say a sophisticated kind of *deus ex machina* reduction base.[18] This is just the sort of familiar but bad reductive reasoning involved in forms of supernaturalism and mysterianism. It amounts to the contention that some suitably unified inscrutable source(s) of power accounts for all observable phenomena. Such reductions do not explain and they do they generate understanding. Second, consider a case where we have a unified global reduction base–for example that proposed by defenders of string theory–the basic concepts of which are arguably of greater internal (semantic) complexity than the disparate phenomena it is supposed to unify. In this particular case, the various phenomena associated with the general theory of relativity with those of quantum mechanics are supposed to be unified by reduction, but the concepts/argument patterns of string theory, although numerically fewer in number, appear to be of much greater mathematical/conceptual complexity than either of the theories it reduces. It is at least plausible then to suggest that we should not count such a reduction as yielding understanding in any serious manner.

[18] Kitcher considers the theological version of this response in his (1980), and concludes that it can be rejected on the basis that it fails to be adequately stringent. However, he offers no substantive account of the stringency of argument patterns. In any case, we can imagine arbitrary non-theological cases that employ sufficiently stringent argument patterns the component terms of which we do not understand at all. Behe's (1996) intelligent design theory might well be a theoretical unification of this sort.

Finally, consider a reduction base that is composed of a small set of inconsistent propositions. From these inconsistent axioms we can trivially derive all other propositions in our belief corpus, but we surely would not want to say that such a set of propositions is coherently understood in some brute manner that explains the propositions derived from them and generates understanding. Such a reduction base cannot even be true, let alone allow us to understand the propositions derived from them.

In the first case we have a reduction base that is numerically small, but which, intuitively, yields no understanding due to its incomprehensibility, and in the second case we find that reduction in the number of factors accepted as primitive in a theory may, ultimately, make for a more conceptually complex (i.e. semantically complex) and less understandable global theory. In the third case we have a reduction in the number of brute facts accepted and a derivation of all other propositions we accept, but no understanding of them because the reduction base is inconsistent. Cases of these sorts can easily be multiplied and articulated, but the details are not really important here. What is important is that cases one and two can be avoided only if the unificationists, by stipulation, simply *equate* understanding with unification in order to obviate S1–S3 and thereby satisfy DE3 trivially as a matter of definition. Given that this is clearly not an acceptable way to justify the steps of the unificationist argument, it seems then, at very least, to be the case that unification is not clearly a sufficient condition for scientific understanding. Cases of the third sort are trickier to avoid and this can be done only by requiring that we eliminate inconsistent reductions from the set of potentially acceptable unifications in some principled manner. As we shall subsequently see, however, this is not realistically possible.

Unification cannot be a necessary condition of explanation for naturalists without leading to outright skepticism, and Kitcher (1981, 1983, 1992, 1993), in particular, is an avowed naturalist.[19] Non-naturalists like Friedman (2000, 2001) are faced with a related, but different, problem. Such, non-naturalists can treat unification as a necessary condition of explanation and avoid skepticism only by attributing *a priori* faculties to us that exceed

[19] Friedman, however, is an avowed non-naturalist as his (2000) and (2001) indicate.

well-verified computational constraints on any feasible inference procedure. How do these particular problems arise? First, consider what the unificationists' view of explanation would require of cognitive agents or cognitive communities. What they must be able to do is to compare total, global, unifications of our knowledge corpus, and then accept the one that is the simplest (in terms of some specified criterion of reductive simplicity) from among the array of possibilities. It turns out that such a task is not computationally satisfiable in realistic times for even finite sets of finite and very small systems of beliefs, even with the aid of all available computational resources.[20] In fact, even if we accept the dubious assumption that the linguistic representations of particular belief corpuses are finite in size and that we need only consider a finite set of such systems, we cannot check the consistency of relatively simple systems with available computational resources in feasible times, let alone check the consistency of and compare *all logically possible* systematizations of our beliefs in terms of their overall simplicity. However, we must be able to check the consistency of such systems or URA might well–in violation of CA1 and CA2–sanction our accepting very simple but inconsistent systematizations of our beliefs as explanatory. As we have seen, this is because all of our beliefs can trivially be derived from inconsistent sets of such axioms. To avoid such cases we must first delimit the (infinite) set of possible alternative systematizations to those that are logically consistent, and then we are supposed to select that consistent systematization from among the remaining set that fares best in terms of URA. However, it is not even *physically possible* to check for the consistency of such systematizations, let alone assess the comparative global simplicity of an infinite set of consistent systematizations of our beliefs.[21]

Given these deeply troubling (but well-known) facts about the computational features of belief systems and URA, unificationists like Kitcher who accept naturalism must be skeptics. This is because we cannot apply URA

[20] See (Kornblith 1989) and (Cherniak 1986) for explicit consideration of such computational restrictions on epistemic processes. Also, see (Harman 1986) for discussion of the computational problem that arise for belief systems of infinite size. It is crucial to note that any belief system closed under logical consequence will be infinite.

[21] (Cherniak 1986) is the canonical source on this point.

in practice in such a way that it can be satisfied and so we cannot really ever explain anything. Given these same facts, unificationists like Friedman, who accept some form of non-naturalism, must attribute to us almost occult *a priori* epistemic faculties that exceed computational/mathematical constraints on feasible, or even physically possible, procedures in order to allow for the satisfaction of URA. Both approaches are obviously unacceptable, and so, at very least, unificationism is either utterly unrealistic or deeply committed to skepticism. Consequently, unification is not a plausible necessary condition for explanation.

4. Two types of understanding

As mentioned in the introduction, it appears to be the case that two senses of ‘understanding’ are often conflated in discussions of scientific explanation. Not surprisingly, the conflation of these two concepts has led to considerable misunderstanding on the part of the parties to the traditional debate concerning scientific explanation and, more recently, to the debate concerning how explanation relates to understanding. However, the ERA provides us with great insight into this error, and this is why it is important that we take the ERA seriously. It just will not do do to sweep it under the rug as traditional and unworthy of serious consideration as the unificationists seem to do.

Ignoring the ERA has made the unificationist view appear to be rather more plausible than it really is. Moreover, this myopia is the root cause of the kinds of problems that afflict the unificationist view raised here. In effect, what unificationists appear to have done by ignoring the ERA is to equate (unwarrantedly and by mere stipulation) scientific understanding with simplifying reduction and simplifying reduction with explanation, thereby trivially satisfying DA3 (i.e. the understanding thesis). In doing so they ignore the concept of *semantic understanding* and this is the concept that is the real core of scientific understanding. Simply put, scientific understanding is not wholly a matter of reductive explanation. Thus, the unificationist view does not satisfy one of the constraints on theories of explanation that the proponents of unificationism themselves accept.

4.1. Reductive understanding and semantic understanding

So, in ignoring the ERA and accepting both the DES and the DMS without justification, the unificationists have overlooked one of the key concepts constitutive of scientific understanding, semantic understanding. This point is related to Heisenberg's observation that,

> For an adequate understanding of the phenomena, the first condition is the introduction of adequate concepts; only with the help of the correct concepts can we really know what has been observed. When we enter a new field, very often new concepts are needed, and these new concepts come up in a rather unclear and underdeveloped form. Later they are modified, sometimes they are almost completely abandoned and replaced by better concepts which then, finally, are clear and well-defined. (Heisenberg 1989, 19)

The relevant point in this passage (that we must have adequate concepts to achieve understanding) is made more poignant when coupled with Heisenberg's view that theoretical formalisms that describe phenomena are distinct from the concepts that, at least in part, are required for understanding of the phenomena in question.[22] For Heisenberg, scientific understanding involves some form of understanding of the concepts that allow us to conceive of the phenomena in terms of some theoretical (i.e. mathematical) formalism. However, from this passage it is clear that Heisenberg believes that understanding comes in degrees and that, often, we adopt a theory without adequate concepts in hand sufficient to generate complete understanding. Ultimately, inadequate and imprecise concepts are replaced by adequate and precise concepts in order to secure scientific understanding of the phenomena in question. We *can* employ a theory in practice without complete understanding of the theory, but then the theory should not be considered to be fully explanatory. It is only when we come to grasp the complete *meanings* of the fundamental terms of a theory adequately, no matter how globally simple the theory is, that we have a *bona fide* explanation of the phenomena it describes.[23]

[22] See (Heisenberg 1930), especially chapters 1-3 and the appendix.

[23] See (Radder 1991; Shaffer 2008a; and 2008b) on this matter.

So, it appears to be the case that a form of semantic understanding is needed here and that semantic understanding involves a form of truth-conditional semantics.[24] As such, a theory of explanation that incorporates this concept seems as if it might do better in accounting for the connection between truth, explanation, and understanding than unificationism does. That the unificationists have overlooked this concept is, again, primarily due to their ignoring the ERA and due to their tacit acceptance of a holistic version of conceptual, or linguistic, role semantics.[25] Given these sorts of views about semantics, the unit of meaning is the complete linguistic or conceptual system and the holistic meaning of such a system is exclusively a function of the logical relations between the constitutive elements of such a system. Both Friedman and Kitcher appear to believe that understanding is a matter of reduction and involves a particular, broadly logical, relation between the sentences that constitute a total theory. They do not appear to accept that understanding involves anything more.[26]

This is a dubious assumption from the perspective of semantics, and it is a plausible way to link explanation to understanding *only* if it is legitimate to ignore the ERA.[27] If we resist this temptation and–recalling our the discussion of DES–recognize that numerical smaller systems of sentences need not be more *semantically* understandable, then we should conclude that unificationism does not, non-trivially and non-skeptically, connect explanation and understanding. Given this observation about the incompleteness of the unificationists' views of understanding, we can then see, at least schematically, what an adequate theory of how explanation relates to scientific understanding will look like.

First, as explanation in general and IBE in particular, must have some truth connection in accord with CA1-CA3, and, as the ERA suggests, we must include *both* semantic understanding (which is truth-conditional) and reductive understanding (which is more holistic) in our account of scientific

[24] Note, however, that there are problems with the computational aspects of some forms of truth-conditional semantics as well. See, for example, (Shaffer 2019).

[25] See, for example, (Harman 1982).

[26] If it does, then they *must* face the ERA with respect to the basic, axiomatic, beliefs in a unified systematization.

[27] See (Fodor and Lepore 1991; and 1992).

understanding. Second, as we are primarily concerned with *human* understanding (which is limited) and should take pains to avoid skepticism about explanation, we must be sure that our theories of both reductive and semantic understanding are *both* computationally tractable.[28] As such, naturalistic study of human computational abilities seems to be a required for the construction of a realistic theory of scientific understanding. Considerations of global simplicity in the sense employed by the unificationists should then be rejected in favor of consideration of local notions of simplicity and this must be conjoined with consideration of the meanings of the basic facts in the reducing theory. So a local body of phenomena is explained by reducing it to a relatively unified but cognitively graspable theory, the basic terms of which are semantically understood. But, science so understood, is a piecemeal operation which often involves integrating local explanatory theories with other previously established explanations. Understanding and explanation are then contextual, come in degrees and are often partial.[29] In real scientific practice, we often only partially grasp and apply a given explanatory theory to a restricted phenomenal domain. Of course, there is much more to be said about what semantic understanding is and how explanatory integration works, but the theory of explanation suggested here promises the possibility of success where unificationism fails.

5. Return to Peirce: local abduction and feasible explanation

There are some clear lessons to be learned from the failure of unificationism, and chief among these are the following results: (a) naturalistic epistemologists ought to reject *global* accounts of explanation, (b) explanation and scientific understanding are, at least in part, fundamentally semantic phenomena, and (c) all realistic accounts of explanation must involve serious consideration of the concept of simplification. A theory of explanation that rejects occult epistemology ought then to be constructed in such a way that it conforms to these desiderata, and it will look quite a bit

[28] See (Shaffer 2019) about some concerns on this point with respect to possible worlds semantics.

[29] See (Shaffer 2012) on the partiality and contextuality of explanation.

more like Peirce's original theory of local abduction than it will look like contemporary unificationist views such as those defended by Friedman and Kitcher. Some foundational work has been done here with respect to such an account, but there is much more work to be done in fleshing out the core concepts of an adequate account of explanation and scientific understanding that reflects these considerations. Amongst the most important tasks then to be undertaken are the tasks of fleshing out of a truth-conditional concept of semantic understanding, the determination of the limits of realistic globalization in our belief systems, and determining how simplification, explanation, and scientific understanding are related with respect to potentially integratable local explanations.

References

Behe, Michael. 1996. *Darwin's Black Box*. New York: Simon and Schuster.

Cartwright, Nancy. 1999. *The Dappled World*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139167093

Cherniak, Christopher. 1986. *Minimal Rationality*. Cambridge: M.I.T. Press.

De Regt, Henk. 2017. *Understanding Scientific Understanding*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780190652913.001.0001

De Regt, Henk, Sabina Leonelli, and Kai Eigner, eds. 2009. *Scientific Understanding*. Pittsburgh: University of Pittsburgh Press.

Dray, William. 1964. *Laws and Explanation in History*. New York: Oxford University Press.

Dupré, John. 1983. "The Disunity of Science." *Mind* 42 (367): 321-46. https://doi.org/10.1093/mind/XCII.367.321

Feigl, Herbert. 1970. "The 'Orthodox' View of Theories: Remarks in Defense as Well as Critique." In *Minnesota Studies in the Philosophy of Science* Vol. 4, edited by Michael Radner and Stephen Winokur, 3-16. Minneapolis: University of Minnesota Press.

Fodor, Jerry. 1974. "Special Sciences, or the Disunity of Science as a Working Hypothesis." *Synthese* 28 (2): 77-115. https://doi.org/10.1007/BF00485230

Fodor, Jerry. 2000. *The Mind Doesn't Work That Way*. Cambridge: M.I.T. Press.

Fodor, Jerry and Ernest Lepore. 1991. "Why Meaning (Probably) Isn't Conceptual Role." *Mind and Language* 6 (4): 328-43. https://doi.org/10.1111/j.1468-0017.1991.tb00260.x

Fodor, Jerry and Ernest Lepore. 1992. *Holism: A Shopper's Guide*. Oxford: Blackwell.

Friedman, Michael. 1974. “Explanation and Understanding.” *The Journal of Philosophy*, 71 (1): 5-19. https://doi.org/10.2307/2024924

Friedman, Michael. 2000. “Transcendental Philosophy and A Priori Knowledge: A Neo-Kantian Perspective.” In *New Essays on the A Priori*, edited by Paul Boghossian and Christopher Peacocke, 367-83. New York: Oxford University Press. https://doi.org/10.1093/0199241279.003.0015

Friedman, Michael. 2001. *The Dynamics of Reason*. Stanford: CSLI.

Harman, Gilbert. 1982. “Conceptual Role Semantics.” *Notre Dame Journal of Formal Logic* 23 (2): 242-56. https://doi.org/10.1305/ndjfl/1093883628

Harman, Gilbert. 1986. *Change in View*. Cambridge: M.I.T. Press.

Heisenberg, Werner. 1930. *The Physical Principles of the Quantum Theory*. Chicago: University of Chicago Press.

Heisenberg, Werner. 1989. *Encounters with Einstein*. Princeton: Princeton University Press.

Hempel, Carl. 1965. *Aspects of Scientific Explanation*. New York: The Free Press.

Hempel, Carl. 1966. *Philosophy of Natural Science*. Englewood Cliffs, N.J.: Prentice-Hall.

Kitcher, Philip. 1976. “Explanation, Conjunction and Unification.” *Journal of Philosophy* 73 (8): 207-12. https://doi.org/10.2307/2025559

Kitcher, Philip. 1980. “A Priori Knowledge.” *The Philosophical Review* 86 (1): 3-23. https://doi.org/10.2307/2184861

Kitcher, Philip. 1981. “Explanatory Unification.” *Philosophy of Science* 48 (4): 507-31. https://doi.org/10.1086/289019

Kitcher, Philip. 1983. *The Nature of Mathematical Knowledge*. New York: Oxford University Press. https://doi.org/10.1093/0195035410.001.0001

Kitcher, Philip. 1989. “Explanatory Unification and the Causal Structure of the World” In *Minnesota Studies in the Philosophy of Science* Vol. 13, edited by Wesley Salmon and Philip Kitcher, 410-505. Minneapolis: University of Minnesota Press.

Kitcher, Philip. 1992. “The Naturalists Return.” *The Philosophical Review* 101 (1): 53-114. https://doi.org/10.2307/2185044

Kitcher, Philip. 1993. *The Advancement of Science*. New York: Oxford University Press. https://doi.org/10.1093/0195096533.001.0001

Kornblith, Hilary. 1989. “The Unattainability of Coherence.” In *The Current State of the Coherence Theory*, edited by John Bender, 207-14. Dordrecht: Kluwer. https://doi.org/10.1007/978-94-009-2360-7_19

Lambert, Karel. 1980. “Explanation and Understanding: An Open Question?” In *Rationality in Science*, edited by Risto Hilpinen, 29-34. Boston: D. Reidel.

Peirce, Charles S. c.1901/1931-1958. *Collected Papers of Charles Sanders Peirce.* C. Hartshorne, edited by Paul Weiss, and Arthur Burks, 8 vols. Cambridge: Harvard University Press.

Radder, Hans. 1991. "Heuristics and the Generalized Correspondence Principle." *The British Journal for the Philosophy of Science* 42 (2): 195-226. https://doi.org/10.1093/bjps/42.2.195

Salmon, Wesley. 1989. "Four Decades of Scientific Explanation." In *Minnesota Studies in the Philosophy of Science* Vol. 13, edited by Wesley Salmon and Philip Kitcher, 3-195. Minneapolis: University of Minnesota Press.

Scriven, Michael. 1970. "Explanations, Predictions and Laws." In *Minnesota Studies in the Philosophy of Science* Vol. 3, edited by Herbert Feigl and Grover Maxwell, 170-230. Minneapolis: University of Minnesota Press.

Shaffer, Michael. 2001. "Bayesian Confirmation of Theories that Incorporate Idealizations. *Philosophy of Science* 68 (1): 36-52. https://doi.org/10.1086/392865

Shaffer, Michael. 2002. "Coherence, Justification, and the AGM Theory of Belief Revision." In *Perspectives on Coherentism*, edited by Yves Bouchard, 139-60. Ontario, Canada: Aylmer-Éditions du Scribe.

Shaffer, Michael. 2008a. "Idealization, Counterfactuals and the Correspondence Principle." In *The Courage of Doing Philosophy: Essays Dedicated to Leszek Nowak*, edited by Jerzy Brzezinski, Anderzj Klawiter, Theo Kuipers, Krzystof Lastowski, Katarzyna Paprzycka, and Piotr Przybysz, 179-204. Amsterdam: Rodopi.

Shaffer, Michael. 2008b. "Re-formulating the Generalized Correspondence Principle: Problems and Prospects." *Polish Journal of Philosophy* 2: 99-115. https://doi.org/10.5840/pjphil2008217

Shaffer, Michael. 2012. *Counterfactuals and Scientific Realism.* New York: Palgrave-MacMillan.

Shaffer, Michael. 2019. "Safety, the Preface Paradox and Possible Worlds Semantics." *Axiomathes* 29 (4): 347-61. https://doi.org/10.1007/s10516-018-9413-3

Woodward, James. 2017. "Scientific Explanation." *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), edited by Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2017/entries/scientific-explanation/

Organon F 27 (2) 2020: 169–186 ISSN 2585-7150 (online)
https://doi.org/10.31577/orgf.2020.27202 ISSN 1335-0668 (print)

RESEARCH ARTICLE

In Defence of the Phenomenological Objection to Mental Fictionalism

Miklós Márton* – János Tőzsér**

Received: 4 May 2019 / Accepted: 15 September 2019

Abstract: In this paper, we defend the main claims of our earlier paper "Mental Fictionalism as an Undermotivated Theory" (in *The Monist*) from Gábor Bács's criticism, which appeared in his "Mental fictionalism and epiphenomenal qualia" (in *Dialectica*). In our earlier paper, we tried to show that mental fictionalism is an undermotivated theory, so there is no good reason to give up the realist approach to the folk psychological discourse. The core of Bács's criticism consists in that our argumentation rests on an equivocation concerning the folk psychological concepts of conscious experiences. In our present argumentation, at first, we shortly recapitulate our earlier argumentation and Bács's main objection to it. After that, we argue against the case of equivocation, claiming that it rests on a highly implausible and unsupported verificationist approach. Lastly, in answering another remark of Bács's, we discuss the possibility of a realist mental fictionalism and conclude that it is an incoherent standpoint.

Keywords: mental fictionalism; folk psychology; mental antirealism; verificationism; conceptual dependence.

* Eötvös Loránd University Budapest
✎ Corresponding author. Center for Theory of Law and Society, Eötvös Loránd University Budapest, Bartók Béla str. 52/B, 2/2, Budapest, 1111, Hungary
✉ marton@ajk.elte.hu

** Hungarian Academy of Sciences
✎ Institute of Philosophy, Hungarian Academy of Sciences, Ráday str. 26, Budapest, 1092 Hungary
✉ jantozser@gmail.com

1. Introduction

In his paper "Mental Fictionalism and Epiphenomenal Qualia" (Bács 2018), Gábor Bács analyzed and criticized our earlier work "Mental Fictionalism as an Undermotivated Theory" in detail. In our earlier paper, we tried to show that mental fictionalism is an undermotivated theory, so there is no good reason to give up the realist approach to the folk psychological discourse.

In the present paper, we reply to Bács's objections. At first, we recapitulate our earlier argumentation shortly. Secondly, we summarize Bács's main objection to it. Thirdly, we argue that this objection fails. Lastly, in answering another remark of Bács's, we discuss the possibility of realist mental fictionalism.

2. What is mental fictionalism and why is it undermotivated?

In our earlier paper, we treated mental fictionalism as a theory about folk psychological discourse. In this sense, it is a *pragmatic* theory: it concerns the *use* of folk psychological sentences, not the content or truth of them. Its core thesis states that when we utter sentences of folk psychology, we do not assert the truth-conditions of the propositional contents of these sentences; that is, we do not use such sentences to describe facts of our mental life, rather we use them for other goals. For example, for evaluating our fellows' behavior or expressing emotions, or making as if we asserted something (see Márton, Tőzsér 2013, 627-28; and Demeter 2013).

To this extent, mental fictionalism is in contrast to those interpretations of the discourse which are committed to the fact-stating nature of the use of folk psychological sentences. Now, we strongly believe that this later realist interpretation is the *default view* of the pragmatics of this discourse. People in their non-philosophical moments take utterances of folk psychological sentences (mental state attributions to ourselves and to our fellows, explanations of their behaviors, etc.) as real fact-stating expressions. We normally think of these sentences as such that people use them with the intention to describe discourse-independent mental phenomena. Since the realist position is the default one, mental fictionalism as an antirealist view must be an *error theory*: it claims that we are usually in error about the pragmatics of the discourse. This is the main point where it differs from the other two antirealist approaches, namely eliminativism and nonfactualism. Though all three can be viewed as an error theory, the other two versions locate the error not in the pragmatic properties of folk psychological sentences, but in their semantic ones. According to eliminativism, we are wrong in taking folk psychological sentences mainly true, and according to nonfactualism we are wrong in taking them to be contentful.

Since mental fictionalism is an error theory, a proponent thereof has to give us a reason why we should not commit ourselves to the realist interpretation of the discourse. She should tell a story *why it is misleading* to see the use of such sentences as stating facts about our real mental life. In short, there must be some motivations to endorse mental fictionalism.

Earlier we identified two conditions for having such a motivation: (a) One can doubt the existence of the entities postulated by folk-psychology, and (b) nevertheless, due to certain (mainly practical) considerations, one does not want to give up this discourse. Now, we argued that the first condition cannot be met. One cannot raise serious doubts about the existence of mental entities postulated by folk-psychology. After all, what else would be misleading in the default realist approach to folk-psychology? If someone had absolutely no doubt about the existence of mental phenomena, why would she want to take the use of its sentences as not intending to describe these phenomena? Especially if she truly does not want to give up the discourse.

Of course, on the surface, mental fictionalism is an ontologically neutral theory, because it deals only with pragmatics, or the use of folk psychological sentences rather than the truth of them. However, one can see now that in order to motivate the choice of this theory, one has to commit herself to the ontological position that the existence of mental entities is, at least, dubious. We think, therefore, that Bács is right when he writes that our objection against mental fictionalism "is an objection to *mental antirealism* in general" (Bács 2018, 302; emphasis in the original). All three antirealist views are committed negatively to the ontological status of mental phenomena, and fictionalism has the weakest form of this commitment. So, when we succeed in proving that it is not tenable, we also show it in the case of the other two stronger positions.

Naturally, the essential point of the question is whether we are right in claiming that the existence of mental entities does not raise serious doubts. In the original paper, we presented two considerations for this thesis. According to the first one, the existence of phenomenally conscious states and events cannot be doubted, because they are constituted by the experienced qualities during these states and events. In other words, conscious experiences are entirely constituted by the way they appear to us. Therefore, because one cannot meaningfully doubt whether what appears to her really does appear to her, one also cannot meaningfully doubt the existence of conscious experiences.

As for the second consideration, conscious experiences are paradigmatic mental entities, or they are even the only ones. We think only they can be called 'mental' in a fundamental and primary sense. As we wrote it:

> According to our natural conviction, if a system or an organism, be it as complicated as you like, does not have any conscious experiences, that is, it does not undergo events that are something it is like for it to undergo, and so the world does not appear to it in any way, then we tend to treat this system or organism as an automat without a mental life. (Márton, Tőzsér 2013, 635)

Furthermore, we argued that other 'mental' entities count as mental only insofar as they bear some appropriate relationship to phenomenally conscious experiential states or events. That is, unconscious mental states (beliefs, non-occurrent desires and hopes, etc.) and processes can be counted as mental insofar as they stand in, for example, an inferential or dispositional relationship to conscious ones.

In sum, our argument for the undermotivation of mental fictionalism has the following logical structure:

(1) The existence of conscious experiences does not raise any difficult ontological questions.

(2) Conscious experiences constitute the totality, or at least the paradigmatic representative core, of mental entities described by folk psychology.

(3) Therefore, the existence of mental entities described by folk psychology does not raise any difficult ontological questions. (1, 2)

(4) Mental fictionalism is motivated if and only if (a) the existence of mental entities described by folk-psychology raises serious ontological questions, and (b) due to certain considerations, we do not want to give up this discourse.

(C) Therefore, the fictionalist interpretation of folk psychology is undermotivated. (3, 4)

3. Bács's main objection

Bács's main case against our argumentation, or as he called it the 'phenomenological objection to mental fictionalism,' consists in claiming that it rests on an equivocation (Bács 2018, 303). He states that the first two premises of the argument cannot be jointly true if we stick to one and the same concept of conscious experience, and conversely: the premises could be equally true just in case the two concepts of conscious experience they contain are different. Since equivocation is not allowed in a sound argument, one of these premises must be false.

According to Bács, the first premise can be true only if by conscious experience we mean event-like phenomenally salient entities. In this interpretation, the content of the concept consists entirely in the spatiotemporal and phenomenal properties of such events: the phenomenal features of a conscious episode determine its type, i.e. whether it is pain, pleasure or an itch, while the spatiotemporal features of it determine which token of that type it is.

However, the second premise can be true if and only if the content of the concept of conscious experience in it contains further ingredients, namely the causal profile of that state. Bács thinks that folk psychological notions of the various conscious experiences entail the typical causal connections these states or events have. To use one of his examples: the folk psychological concept of itching contains in itself the allusion to the fact that by those who have this experience, itching causes scratching or at least the urge to do so (Bács 2018, 305). So, folk psychological concepts of conscious experiences are not exhausted by their spatiotemporal and phenomenal properties, but "are conceptually linked to their causes and effects" (Bács 2018, 303)—this is what Bács calls the Conceptual Dependency Thesis (CDT).

His main argument for CDT depends on our practice of attributing conscious mental states and events to our fellows.[1] We routinely do this and of course, in doing so we lean on observed behavior and circumstances as evidence. So far, so good. However, Bács goes further when he maintains that

> there is no conceptually innocent, purely empirical evidence. Any evidence must be plugged into the concept of the thing it is evidence for. This should not come as a surprise, because evidence is an epistemic notion. What counts as evidence for X depends in part on what we know about X, and what we know about X is rolled into its concept. A necessary condition for being evidence for is this: Y is evidence for X only if X is conceived as (i.e. the concept of X means) such a thing that under normal circumstances if Y obtains it stands a good chance that X obtains; or alternatively, if it is not possible (or highly improbable, or whatever) that Y obtains but X fails to obtain. […] So the concept of X must include the connection between X and Y, which constitutes the conceptual link, in order for Y to be able to count as evidence for X, and therefore, it will include Y. (Bács 2018, 304)

In short, Bács's reason for CDT is epistemological. He seems to think that the content of a predicate incorporates the way we get to know whether the predicate is instantiated in a certain context.

Let us turn to the first premise of our argument, i.e. the claim that the existence of conscious experiences does not raise any difficult ontological questions and see whether it can be true if we understand the notion of 'conscious experience' as involving CDT. Bács writes that "if conscious

[1] There is a further, minor argument that aims to support CDT in Bács's paper which alludes to linguistic evidence. There are a lot of expressions in natural language which refer to conscious experiences—sensations, perceptual states—which name the part of the body where the causes of these conscious states occurred, or the typical behavior caused by the states. However, as Bács himself hastily adds, this linguistic evidence is very weak, since "it is not always a good idea to take the meaning of ordinary expressions at face value" (Bács 2018, 304). We concur in this question: one should not draw metaphysical conclusions from the meanings of natural linguistic expressions, unless one also wants to achieve serious astronomical insights from the expression "the Sun comes up."

experiences are not entirely constituted by qualia [but also by causal connections], then the indubitable existence of qualia does not imply the indubitable existence of conscious experiences" (Bács 2018, 306). Causal relations are—at least according to most views of causation—empirical and are not, contrary to their phenomenological features, exhausted by their appearance to a subject. For this reason, one can be easily wrong about their existence. So, if a conscious experience is present just in case its usual causal connections are also present, then the existence of conscious experiences is by no means indubitable.

Bács supports this claim with a thought-experiment of the usual Twin-Earth kind. His Twin-Earth is a Leibnizian one, where conscious experiences of the inhabitants of this planet are causally totally isolated. They do not have any causal connections, including the ones our phenomenally identical counterpart conscious experiences have. However, notwithstanding this situation, the inhabitants have folk psychology which is exactly the same as ours. In their folk psychology there are mental predicates standing for conscious experiences and these include the allusion to causal connections like ours—at least according to Bács. From these premises Bács concludes that Twin-Earthers cannot make true assertions by these predicates, since they stand for nothing, as nothing satisfies the description contented in them.

In sum, the alleged equivocation consists in the fact that the first premise, if true, contains a notion of conscious experience in which there are entirely transparent events or states—which is why we cannot be wrong about their existence—while the second premise, if true, contains the folk psychological notion of conscious experience, which, in turn, is committed to CDT. You cannot substitute the two notions in the two premises *salva veritate*, therefore they are different concepts and the case of equivocation is sound.

4. Objections to Bács's objections

We think there is no equivocation in our argument; at least Bács's objection does not prove that there is. Let us start with his first consideration, namely that folk psychological concepts of conscious experience would involve CDT.

As we saw, his main reason for this thesis consists in the assumption that the evidence we lean on when we attribute conscious experiences to our fellows are plugged into the very concepts we have of them. That is—as it can be seen from the above longer quotation—he supposes that the epistemic conditions of true attributions are built into the *meaning* of the predicates we use to describe these conscious episodes. This move has a strong verificationist flavor—at least to our ears. The case of verificationism is supported by the fact that somewhat later (Bács 2018, 305), Bács alludes to the learning of folk psychological notions of conscious experiences in order to justify CDT, and there he obviously supposes that the meaning of a concept entails the circumstances of learning it.

Now, as it is well-known, verificationist theories of meaning have to face some powerful objections and therefore have not been too popular among meaning-theorists over the last sixty years. Intuitively, you need not know how to recognize whether a predicate is instantiated in order to understand it, and even if you know this, this knowledge is not built into the meaning of the predicate. After all, we have not a faint idea how to get to know whether a subatomic particle is present in a given context, but nevertheless, we think we understand the notion of an electron. Or, we strongly believe we do know what dinosaurs are, although we cannot decide whether some fossils serve as good evidence for their existence or not. But even for the paleontologist who surely possesses this knowledge, the concept of a dinosaur hardly entails allusion to the fossils. When she speaks about dinosaurs, it seems she speaks about animals that lived on Earth more than 65 million years ago, and not about present-day fossils.

Moreover, Bács manifestly commits himself to applying the verificationist theory to folk-psychology. For example, in presenting his argument, he writes that "the folk psychological concept of any conscious experience will include the behavior characteristic to it as evidence for its presence in others" (Bács 2018, 304). He also claims that "the folk psychological concept of pain is associated not just with pain sensation, but also with bodily damage causing it and pain behavior it causes" (Bács 2018, 303). What is more, he explicitly states that his CDT is analogous to the so-called criterial solution to the problem of other minds, which in turn "also implies that in folk psychology mental events are not only causally linked to behavior but conceptually as well" (Bács 2018, 305). Criterial evidence is conventional in the sense that it is part of the meaning of a term such as 'pain;' that certain kinds of behavior count as more or less defeasible evidence for its ascription. The criterial solution is therefore also committed to the verificationist theory of meaning to some extent, and therefore, it renders the problem of other minds—which posits a gap between the meaning of conscious state attributions and the justifications thereof—meaningless rather than solving it.

As it is also well-known, there are many persuasive counter-arguments against such views that take behavioral evidences as built into the contents of folk psychological concepts. The most relevant one in our case is Hilary Putnam's famous example of "super-stoics:"

> Imagine a community of 'super-spartans' or 'superstoics'—a community in which the adults have the ability to successfully suppress *all* involuntary pain behavior. They may, on occasion, admit that they feel pain, but always in pleasant well-modulated voices [...] However, they do feel pain, and they dislike it (just as we do). [... I]magine a world in which there are not even pain *reports.* I will call this world the 'X-world'. In the X-world we have to deal with 'super-super-spartans'. These have been super-spartans for so long, that they have begun to suppress even *talk* of pain. [...] They pretend not to know either the word or the phenomenon to which it refers. [...] Only, of course, they do have pains, and they know perfectly well that they have pains. (Putnam 1965/2002, 49-50)

What these examples show persuasively in our opinion is that there could be cases in which the usual causal connections between conscious experience (e.g. pain) and its behavioral effects do not hold. And this situation can be normal in a community or the whole world, so one cannot think that the causal connections in question are usual. So, the folk psychological sentence "X is in pain and X does not show any pain-behavior" does not contain any logical or semantic contradiction. Therefore, the folk psychological concepts of conscious experiences have no conceptual connection to pain-behaviors, contrary to what Bács assumes. The same can be said about the criterial solution to the problem of other minds—at least according to most researchers of the problem. One can coherently conceive of situations where the allegedly criterial evidence holds, but the conscious episode does not occur, or vice versa. The criterial evidence, therefore, has no conceptual link to conscious episodes. There is a gap between observed behavior and the unobserved inner states they fail to bridge (see Hyslop 1995, ch. 5. and 8).

In summary, Bács's arguments in favor of CDT fails to justify the thesis. Naturally, this result does not imply the falsity of CDT—it can be true for other reasons. For example, one can hold such a claim as a result of a functionalist conviction. However, there are well-known counter-arguments against it which Bács fails to account for. For example, the much-debated arguments based on the inverted spectrum thesis or inverted Earth scenarios (whether they are real or counterfactual ones) equally aim to prove that conscious mental states and their normal causal inputs and outputs do not stand in a tight conceptual relationship (see Shoemaker 1982; Block 1990). One can have a red-seeing experience triggered by green objects and followed by events and states usually follow green-seeing experiences. Or, to use the above-mentioned example of Bács: one can have an itching experience without feeling any urge to scratch oneself, but rather to do something else. The proponents of the argument see no conceptual incoherence in such scenarios, even in the case of a whole community. Again, by invoking these arguments, we do not want to claim that CDT is definitely false but merely that they represent a strong challenge to the thesis, so these arguments must be considered by everyone who tends to accept or reject the thesis in question.

Based on the above considerations, we can safely state that Bács failed to justify CDT, that is, the claim that folk psychological notions of conscious mental states conceptually involve and therefore entail their causes and effects. Because of this failure, Bács did not succeed in showing that these notions of folk-psychology cannot refer to conscious experiences as those phenomenally transparent, event-like entities we assume them to be.

As for his considerations concerning the first premise of our argument, we think it is an obvious truth that it cannot be right if we read the concept of conscious experience as involving CDT. Bács is right in claiming that if the allusion to causal connections is plugged into the concepts of conscious experience, the existence thereof is by no means indubitable. We think it is common sense, and therefore we completely agree with Bács about this. Clearly, we have not been persuaded that the folk psychological notion of conscious experience really involve CDT, so nothing seem to threatens the truth of our first premise.

However, we think his Twin-Earth scenario is somewhat misleading, because it contains a highly implausible assumption and therefore, his argument based thereon is not persuasive either. Our problem is the following: Bács's scenario presupposes that the introduction and development of Twin-Earth folk-psychology has run its course entirely *independently* of the phenomenal states the Twin-Earthers had. As if the semantic properties of their folk psychological predicates would have nothing to do with the phenomenal states they undergo. This is the feature of the thought experiment which seems most implausible to us. We cannot believe that the introduction of these predicates was by pure stipulation. Rather, it is much more plausible that the intention of the first users of these terms aimed to name the very phenomenal features they experienced. These experiences were there and were salient at the time of their introductory use. We think that various kinds of conscious experiences are natural kinds, so the predicates in question are natural kind terms. Therefore, if one accepts this plausible reading of the scenario, one has to conclude that these terms *do refer* to the phenomenal states Twin-Earthers undergo. Of course, Twin-Earthers *are wrong* in thinking about these states as having causal connections. They have false beliefs about the nature of their experiences, but they can talk about them successfully.[2] In short, we think that in the plausible reading of the scenario Bács is wrong when he states that "there are no *pains* on Twin-Earth" (Bács 2018, 306; *pain* is the Twin-Earthian folk psychological counterpart concept to our concept of PAIN). There are indeed *pains* on Twin-Earth, namely the conscious experience to which this concept refers, that is, the experience which was present at the time the term was introduced. For the same reason we think Bács is wrong when he writes the following:

[2] We think our reading of the scenario is the one which is in harmony with the original point of Putnam's Twin-Earth example and the argument based on it. See (Putnam 1975; and also Kripke 1972/1980).

> So, just as in the standard case of 'Water' which fails to refer to the liquid found on Twin-Earth because it is not H_2O (Putnam 1975), 'Pain' would fail to refer to what feels like pain on Twin-Earth because it is causally not responsible for pain behavior. (Bács 2018, 308)

In this passage, it remains ambiguous which language the term 'pain' belongs to. If we mean it in harmony with Bács's scenario, it must belong to Twin-Earther's language. In that case, it is analogous to the meaning of the term 'water' also of the language of Twin-Earthers. And the Twin-Earthian term 'water' surely refers to Twin-Earthian water, even if they would falsely think it is composed of two hydrogen and one oxygen atoms. The same is true for the Twin-Earthian term (and concept) of 'pain.'[3]

[3] There is a possible complication here concerning the epiphenomenal nature of Twin-Earthian conscious experiences. Someone might object to the story we described that because these phenomenal states are causally impotent, they cannot cause the introduction of any term or the intention to introduce one. However, we think if it is indeed a problem, it is not just our problem; rather it is also a problem for Bács. At the end of his paper, he presents some supposedly false propositions from Twin Earth folk-psychology which attribute causal connections to the phenomenological states the inhabitants of this planet have. The only natural reading of these propositions, we think, is the one in which by uttering or thinking these propositions, Twin-Earthers speak or think about their phenomenological states. For example, one of Bács's example is the proposition: "Peter did not go into the water because he was *afraid* of sharks" (Bács 2018, 307). We agree with him that this proposition is false, because the phenomenal state Peter has has no causal effect, so it cannot cause his reluctance to go into the water. However, as Bács himself acknowledges, "Peter did feel something which was phenomenologically like the feeling of fear" (Bács 2018, 307), and the natural reading of the sentence is the one in which Peter has false beliefs about this feeling. But then, it seems the only plausible explanation of this fact commits him to the view that this phenomenally salient state caused Peter's thought. Even in the case of Bács's last example, "There exists *anxiety*" (Bács 2018, 308), the natural reading is the same again, namely that the utterer of this sentence (probably a Twin-Earthian philosopher) was speaking about her phenomenal state, just wrongly subsumed it under the concept *ANXIETY*, because it has no causal connections. However, it sounds highly paradoxical if we consider how she can get this thought. If her phenomenal state is indeed epiphenomenal, how can she think about it? In conclusion, if the epiphenomenal nature of Twin

In summary, we agree with Bács that one can be wrong about the existence of causal connections and, therefore, the first premise of our argument cannot be right if we read the concept of conscious experience as involving CDT. Only we think the argument he presents for this statement is not persuasive. All in all, we think we have shown that Bács's case of equivocation is not sound: the two premises can be jointly true with the same reading of 'conscious experience,' namely the one which describes them as simple spatiotemporal entities with phenomenal properties. Therefore, his counter-argument against our phenomenological objection fails.

5. Is there a really realist fictionalist position?

There is a further possible objection to our position, i.e. that the acceptance of mental fictionalism is undermotivated. It is based on the possibility of a *realist fictionalist* position, which, while acknowledging the existence of propositional attitudes, claims that the folk psychological explanations containing these attitudes are fundamentally flawed, and this feature of the discourse would motivate the fictionalist approach. This objection is presented in a somewhat sketchy way by Bács and was also considered by us in our original paper. As he puts it:

> Folk psychological explanations in terms of propositional attitude attributions are fundamentally flawed not because propositional attitudes do not exist, but because propositional attitudes cannot meet important conceptual and methodological requirements for figuring in explanations—for example, because propositional attitudes are individuated by the very behaviors they are meant to explain, or because propositional attitudes are attributable on normative grounds rather than empirical facts. But we cannot

Earth phenomenal states makes our scenario inconsistent, it will make Bács's scenario inconsistent *as well*, since otherwise how can the inhabitants of this strange planet think or speak about their phenomenal states, even if wrongly? We think what this problem really shows is how problematic it is to conceive of epiphenomenal conscious experiences coherently. For more on this problem see (Shoemaker 1975; Chalmers 1996, 172-209).

> hope to purge propositional-attitude discourse from everyday life. Therefore, our best option is to go fictionalist. […] This would be the *realist fictionalist's way*. The realist fictionalist is someone who takes some discourse as a fiction without disputing the existence of its subject matter. (Bács 2018, 302; emphasis in the original)

Although in our earlier paper we considered this objection as a real and serious challenge, we changed our minds by now. We cannot see realist fictionalism as a real theoretical possibility.

Let us start by asking the question: what could be one's reason to acknowledge the existence of propositional attitudes? We think there are two possible answers: a) one could have only experiential, i.e. explanation-independent reasons to believe in the existence of propositional attitudes; and b) the only reason at hand could be following Quine's dictum, namely that one has to believe in the existence of only those entities which play a role in successful theoretical explanations or explanatory strategies. So, according to the first option, propositional attitudes are the kinds of entities that are phenomenally salient or, at least, appropriately related to phenomenally salient mental episodes. On the other hand, according to the second option, propositional attitudes are entities of a theoretical kind. In the first case, one has reason to believe in the existence of propositional attitudes if and only if one has appropriate experiences, while in the second case, if and only if there are successful theoretical explanations at hand in which propositional attitudes play a role.

Applying these two options, we can delineate the logical landscape of possible positions. There are four theoretical positions according to how one answers the relevant questions in the two options. By the first option, this question asks whether there are appropriate *experiences* of propositional attitudes, while in the second option, the question asks whether there is a successful *explanation* which appeals to them. Let us see these possibilities and what follows from them concerning the alleged realist fictionalist position.

If you choose the first option, you have to decide whether you think there really are explanation-independent experiential reasons to acknowledge the existence of propositional attitudes. If your answer is positive, then you will have good experiential reasons to believe in propositional attitudes. So, you are surely a *realist* about them. Furthermore, you do not want to give up folk psychological discourse *tout court*. So, why be a *fictionalist*? Naturally, you can think that there are some bad explanations in folk-psychology (they may have weak explanatory power, falsified predictions or results etc.), and indeed, you may be right about this. However, in such a situation, you would be motivated to improve folk-psychology and not to choose fictionalism. That is, you would try to present better explanations, more exact predictions, etc. using the *same* propositional attitude terms. Indeed, current psychology seems to do exactly this. It seems very implausible to us to think that in this situation you would be inclined to take the usage of sentences containing propositional attitudes as non-fact-stating ones. In sum, in the case of this position, you will have no reason to opt for realist fictionalism, because you will have no reason to choose fictionalism at all.

If you think there are no good experiential reasons for accepting the existence of propositional attitudes, or you think it is dubious whether there are, then you will have indeed *ontological doubts* about the existence of propositional attitudes. So, you claim that propositional attitudes are phenomenally salient, experiential kinds of entities (or, at least, are appropriately connected to such kind of entities), for the existence of which there are no theoretical reasons, but you think there are no experiential reasons, either. Therefore, you are not a realist about propositional attitudes, *a fortiori* you cannot be a *realist* mental fictionalist. It is that simple.

Let us now turn to the other main option, i.e. the one which follows Quine's dictum. Of course, you have to choose again whether you think these explanations or explanatory strategies succeed or not.

If your answer is positive, then we will think again that there is no reason to be a realist fictionalist. The situation is very similar to (1): you have good reason to believe in the existence of propositional attitudes and you do not want to give up folk-psychology. Moreover, in this case you think that—at least the majority of—folk psychological explanations are good ones. So, apparently you have every reason to be a realist about the existence of propositional attitudes, but you have absolutely no motivation to accept the *fictionalist* approach. The fact that the only reason to acknowledge propositional attitudes is that they play a role in successful folk psychological explanations does not offer any motivation for giving up the default realist interpretation of the pragmatics of this discourse. You can say that it motivates you to take propositional attitudes as theoretical entities, but why should we not use sentences containing terms of theoretical entities to state facts? For example, one can plausibly argue that genes are explanation-dependent theoretical entities, but it would be absurd to conclude from this that evolutionary biologists do not use sentences containing the term 'gene' to state real facts.

A good illustration of this position would be that of Dennett's. He famously holds that we have to posit propositional attitudes only because of the success of the intentional strategy in explaining the behavior of our fellows. And, at the same time, he explicitly denies to be a mental fictionalist. As he puts it:

> Some instrumentalists have endorsed *fictionalism*, the view that certain theoretical statements are *useful falsehoods*, and others have maintained that the theoretical claims in question *were neither true nor false* but mere instruments of calculation. *I defend neither of these varieties* of instrumentalism; as I said when first I used the term above: "people really do have beliefs and desires, on my version of folk psychology, just as they really have centers of gravity." (Dennett 1987b, 72; the first two emphases are in the original, the third one is ours)

Of course, there are cases where folk psychological explanations, or, as he calls it, "the intentional stance" does not work because of the failure of the assumption of rationality. However, it is crucial that in these cases Dennett *does not want to maintain* the intentional, i.e. folk psychological discourse. As he puts it: "This is not to say that we are always rational, but that when we are not, the cases *defy description* in ordinary terms of belief and desire" (Dennett 1987a, 87; emphasis is ours). In such situations, we have to step back and apply other kinds of explanations.

The last possibility is the one where you think that there could be only explanation-dependent reasons to acknowledge propositional attitudes, but you also think that these explanations are flawed. In this situation, the only rational conclusion, we think, is that there are no propositional attitudes. What else could you think? If you take the explanations in question as unsuccessful and, at the same time, you also think that propositional attitudes are theoretical entities whose existence depends on the success of the explanations they take part in, then you will have no other logical option but to deny the existence thereof. So, if that would be the reason to be a fictionalist, it is indeed based on *ontological doubts*. Consequently, you cannot be a realist about propositional attitudes, *a fortiori*, you cannot reasonably choose *realist* mental fictionalism. The presence of ontological doubts is explicitly stated by a committed fictionalist, namely Demeter in the following way:

> Folk psychology represents agents in a way similar to how some fictions represent the world: in a way they are not, and—as folk psychology does not state facts—cannot be. In this sense folk psychology is a tool for making Escherian representations. Escher's Drawing Hands, for example, is not a representation of hands drawing one another, but a representation as if hands were drawing one another—as if it were possible. (Demeter 2013, 497)

In other words, according to Demeter, folk psychological sentences are about impossible states of affairs, so they cannot exist. What is this if not a radical ontological doubt about propositional attitudes?

To conclude, Bács's objections did not persuade us that our earlier argumentation against mental fictionalism was wrong. We still think our "phenomenological objection" does show that mental fictionalism is undermotivated. Moreover, we think that the realist fictionalist approach mentioned by Bács is not even a consistent standpoint, therefore it can hardly motivate the acceptance of mental fictionalism, too.

Acknowledgements

We are grateful to Zsuzsanna Balogh for her helpful comments on an earlier version of this paper. We also are grateful for the anonymous referee for her/his helpful comments and suggestions.

Funding

The research leading to this paper was supported by NKFIH (National Research, Development and Innovation Office), grant no. K 123839.

References

Bács, Gábor. 2018. "Mental Fictionalism and Epiphenomenal Qualia." *Dialectica* 72 (2): 297–308. https://doi.org/10.1111/1746-8361.12229

Block, Ned. 1990. "Inverted Earth." *Philosophical Perspectives* 4: 53–79. https://doi.org/10.2307/2214187

Chalmers, David. 1996. *The Conscious Mind.* Oxford: Oxford University Press.

Demeter, Tamás. 2013. "Mental Fictionalism: The Very Idea." *The Monist* 96 (4): 483–504. https://doi.org/10.5840/monist201396422

Dennett, Daniel C. 1987a. "Making Sense of Ourselves." In *The Intentional Stance*, 83–101. Cambridge Mass.: MIT Press.

Dennett, Daniel C. 1987b. "Reflections: Instrumentalism Reconsidered." In *The Intentional Stance*, 69–81. Cambridge Mass.: MIT Press.

Hyslop, Alec. 1995. *Other Minds.* Dordrecht: Kluwer.

Kripke, Saul. 1972/1980. *Naming and Necessity.* Cambridge Mass.: Harvard University Press.

Márton Miklós, and Tőzsér János. 2013. "Mental Fictionalism as an Undermotivated Theory." *The Monist* 96 (4): 622–638. https://doi.org/10.5840/monist201396429

Putnam, Hilary. 1965/2002. "Brains and Behavior." In *Philosophy of Mind: Classical and Contemporary Readings*, edited by David Chalmers, 45–54. Oxford: Oxford University Press.

Putnam, Hilary. 1975. "The Meaning of 'Meaning'." *Minnesota Studies in the Philosophy of Science* 7: 131–193.

Shoemaker, Sydney. 1975. "Functionalism and Qualia." *Philosophical Studies* 27 (5): 291–315. https://doi.org/10.1007/BF01225748

Shoemaker, Sydney. 1982. "The Inverted Spectrum." *The Journal of Philosophy* 79 (7): 357–381. https://doi.org/10.2307/2026213

Organon F 27 (2) 2020: 187–262
https://doi.org/10.31577/orgf.2020.27203

ISSN 2585-7150 (online)
ISSN 1335-0668 (print)

RESEARCH ARTICLE

Boulesic Logic, Deontic Logic and the Structure of a Perfectly Rational Will

Daniel Rönnedal

Received: 22 January 2019 / Accepted: 4 November 2019

Abstract: In this paper, I will discuss boulesic and deontic logic and the relationship between these branches of logic. By 'boulesic logic,' or 'the logic of the will,' I mean a new kind of logic that deals with 'boulesic' concepts, expressions, sentences, arguments and systems. I will concentrate on two types of boulesic expression: 'individual x wants it to be the case that' and 'individual x accepts that it is the case that.' These expressions will be symbolised by two sentential operators that take individuals and sentences as arguments and give sentences as values. Deontic logic is a relatively well-established branch of logic. It deals with normative concepts, sentences, arguments and systems. In this paper, I will show how deontic logic can be grounded in boulesic logic. I will develop a set of semantic tableau systems that include boulesic and alethic operators, possibilist quantifiers and the identity predicate; I will then show how these systems can be augmented by a set of deontic operators. I use a kind of possible world semantics to explain the intended meaning of our formal systems. Intuitively, we can think of our semantics as a description of the structure of a perfectly rational will. I mention some interesting theorems that can be proved in our systems, including some versions

* Stockholm University
✎ Department of Philosophy, Stockholm University, Universitetsvägen 10 D, 106 91 Stockholm, Sweden
✉ daniel.ronnedal@philosophy.su.se

of the so-called *hypothetical imperative*. Finally, I show that all systems that are described in this paper are sound and complete with respect to their semantics.

Keywords: Boulesic logic; deontic logic; modal logic; practical rationality; the hypothetical imperative; the logic of the will; semantic tableaux.

1. Introduction

In this paper, I will discuss boulesic and deontic logic and the relationship between these branches of logic. By 'boulesic logic' (from the Greek 'boulesis'), 'the logic of the will,' 'conative logic,' or 'volative logic' I mean a kind of logic that deals with 'boulesic' concepts, words, expressions, sentences, principles, arguments and systems. Boulesic logic is a new kind of logic. There are hints about such a logic in the literature, but few attempts to develop the basic idea in detail. The main results in this paper are therefore entirely new.[1]

[1] In the *Nicomachean Ethics* (Book VI and VII), Aristotle mentions a special kind of practical syllogism where the conclusion is a command or act. This suggests that the Greek philosopher might have envisioned some sort of 'practical logic.' Immanuel Kant discusses some principles that it is plausible to be included in a boulesic logic, for instance, the so-called *hypothetical imperative* (see below for more on this). Some similar principles were discussed already by various medieval thinkers (see Knuuttila 2004, 3.3). In 1926, Ernst Mally published the book *Grundgesetze des Sollens*, which is generally thought to contain the first published formal deontic system ever. Mally's book has the subtitle *Elemente der Logik des Willens* (Elements of the Logic of (the) Will), which indicates that he saw important connections between deontic logic and the logic of the will. In fact, he might have thought they are the same thing. There are some similarities between boulesic logic and 'intentional logic,' even though there are also many important differences. Later in the introduction, I will say more about this. For some information on intentional logic, see, for example, (Broersen 2011; Broersen, Dastani and van der Torre 2001; Cohen and Levesque 1990; Lorini and Herzig 2008; Marra and Klein 2015; and Semmling and Wansing 2008). Harry Gensler develops a logic of the will as a part of a kind of 'belief logic' that is based on imperative logic and extends ideas introduced by Hector-Neri Castañeda in several works (see Gensler 2002, Chapter 10 for more on this). See also (Bratman 1999).

Some examples of boulesic words include ‘wanting,’ ‘willing,’ ‘accepting,’ and ‘consenting.’ Many words and expressions in the vicinity might also be classified as boulesic, such as ‘intending,’ ‘desiring,’ ‘rejecting,’ ‘loving,’ and ‘hating,’ and—more generally—‘having a pro-attitude.’ A boulesic concept is a concept expressed by a boulesic word. A boulesic sentence is a sentence that (essentially) includes a boulesic word. A boulesic argument is an argument that (essentially) involves a boulesic sentence, and a boulesic system is a system that (essentially) includes various boulesic axioms and/or rules of inference.[2]

Here are some examples of boulesic sentences:

- John wants to win.
- Jennifer wants it to be the case that there will be peace.
- Sonny accepts the fact that he is never going to be a professional football player.

Here are some examples of boulesic principles (not necessarily valid):

- No one wants it to be the case that A and also wants it to be the case that not-A.
- If a perfectly rational individual x wants it to be the case that A and B is a necessary means to A, then x wants it to be the case that B.
- It is permitted that you perform this action only if everyone who is perfectly rational consents to the idea that you perform this action.

Here are some examples of boulesic arguments (not necessarily valid):

Boulesic logic, in the sense that I am using the term, is similar to deontic logic, intentional logic, imperative logic (if there is such a thing) and certain forms of epistemic and doxastic logics, and there seem to be important connections between, for instance, boulesic and deontic logic (for more on this, see sections 3.4, 3.5 and 4.3). However, there are also important differences between these branches, and there are, as far as I know, no systems of the kind introduced in this paper in the literature.

[2] These are not meant to be exact definitions. There are sentences that include boulesic words that are not boulesic—for example, ‘The word “want” is an English word’ and ‘Jim believes that Greta wants to become a doctor’—and there are arguments that include boulesic sentences that are not boulesic, etc. Hence, I have added the qualification ‘essentially.’

Argument 1

1. Henrietta wants the sun to shine.
2. Henrietta wants to go to the beach.
   Hence:
3. Henrietta wants the sun to shine and go to the beach. [From 1 and 2]

Argument 2

1. Mona wants to be happy.
   Thus:
2. If Mona is perfectly rational, she consents to the idea that she is happy. [From 1]

Argument 3

1. Every person in the class wants to pass the exam.
2. Sandra is a person in the class.
3. It is necessary that Sandra passes the exam only if she studies hard.
   Therefore:
4. If Sandra is perfectly rational, she wants to study hard. [From 1—3]

I will concentrate on two types of boulesic expression in this paper: 'individual c wants it to be the case that A' and 'individual c accepts that it is the case that A' or 'individual c consents to the state of affairs (the fact/the idea) that A.' These expressions will be symbolised by two sentential operators, $\mathcal{W}$ and $\mathcal{A}$ respectively, which take individual terms and sentences as arguments and give sentences as values. That is, in the formal sentence $\mathcal{W}_cB$, the constant c refers to an individual and B is any well-formed sentence (and similarly for $\mathcal{A}_cB$). In other words, '$\mathcal{W}_cB$' is read as 'c wants it to be the case that B' and '$\mathcal{A}_cB$' is read as 'c accepts that it is the case that B.'

Deontic logic is a relatively well-established branch of logic. It deals with normative concepts, sentences, arguments and systems. For introductions to this branch of logic, see for example (Åqvist 1987, 2002; Gabbay et al. 2013; and Hilpinen 1971, 1981). In this paper, I will show how deontic logic can be grounded in boulesic logic in a certain sense. I will develop a set of semantic tableau systems that include boulesic and alethic operators, possibilist quantifiers and the identity predicate, and I will show how these systems can be augmented by a set of deontic operators. I use a kind of possible world semantics to explain the intended meaning of our formal systems. Intuitively, we can think of our semantics as a description of the structure of a perfectly rational will. I mention some interesting theorems that can be proved in our systems, including several versions of the so-called *hypothetical imperative* (if you want A and B is a necessary means to or condition for A, then you also want B insofar as you are rational; more about this below). Finally, I show that all systems that are described in this paper are sound and complete with respect to their semantics.

$\mathcal{W}$ takes any sentence as its argument. In $\mathcal{W}_cB$, B can be any well-formed formula whatsoever. So, it is possible to want anything, so to speak. B can be about the present time (I want you to be here now), about the future (She wants to become a doctor [some time in the future]) or about the past (I want [hope, desire, wish] that I made the right choice yesterday (Feldman 2004, 62)); it can be about a contingent state of affairs (She wants to buy a house) or a necessary state of affairs (He wants this mathematical theorem to be true); it can be about facts concerning nature (He wants the sun to shine tomorrow) or about various mental states (I want to feel happy); it can be about c (He wants to be happy) or about some other individual or individuals (She wants her children to be happy), and so on. According to our systems, it is meaningful to speak about wanting anything, and it is (logically) possible that any sentence of the form $\mathcal{W}_cB$ is true. It is even possible for someone to want something that is impossible. It seems reasonable to me that our systems allow this. Normally, if we want B, it is probably true that B is a contingent state of affairs that is in (or about) the future. However, this does not appear to be logically necessary. The same is true about $\mathcal{A}$ (acceptance). If we assume that c is perfectly rational (reasonable or wise), things are different. It seems plausible, for example, to claim that every perfectly rational (reasonable or wise) individual only wants something if it is possible; let us call this principle the Want-Can principle ($\mathcal{W}C$).[3] Nonetheless, there is a difference between what a perfectly

[3] In every system that includes the tableau rule $T-\mathcal{W}C$ (Table 18), we can prove this principle. $T-\mathcal{W}C$ is valid in the class of all models that satisfy $C-\mathcal{W}C$ (see Section 3.3.5 for more on this). Note that the Want-Can principle does not entail

rational individual wants and what some arbitrary agent desires. It appears to be possible for someone who is not perfectly rational to want something that is impossible. In fact, there are probably actual examples of people with an inconsistent will. Yet, this should come as no surprise: not everyone is perfectly rational.[4]

When we say that someone wants something (or accepts something), we usually mean that she wants (accepts) this 'thing' in an *all-things-considered* sense in this paper; we do not necessarily mean that she wants it *in itself*. It is possible to want something as a means to something else and it is possible to want something in itself and it is possible to want something all-things-considered. Someone can, for example, want to study for the exam in an all-things-considered sense even though she does not want this in itself. She wants to study for the exam because she wants to pass the exam and she believes that she will pass the exam only if she studies for it. Studying is a means to an end. Moreover, it is possible for someone to want (or accept) A in an all-things-considered sense even if she does not like every aspect of A or every consequence of A and even if she has some desire (a prima facie desire) for not-A.

that everyone ought to want something only if it is possible. The latter thesis is independent of the Want-Can principle. The Want-Can principle is even compatible with the proposition that some individuals that are not perfectly rational ought to want some things that are impossible (in every system in this paper). Furthermore, it may still be reasonable to think about doing something impossible, to daydream about doing something impossible, and perhaps also to wish that something impossible be the case. But (merely) thinking, wishing, daydreaming, and so on is not the same thing as wanting.

[4] Some philosophers seem to think that desires (and wants) are always future-oriented—that is, they think that if at time t, someone S desires that p be the case, then p is future relative to t. Wayne Sumner might be an example (Sumner 1996, 128–30; Sumner 2000). Other philosophers appear to reject this thesis (see, for example, Feldman 2004, 61–63). According to our systems, wants are not necessarily concerned with the future; it is possible that they are directed at the present or the past too, for instance. Still, I am inclined to believe that all 'genuine' wants are future-oriented for *perfectly rational individuals*—at least if we assume that facts about the past and the present are historically settled. Thus, I think our systems can do justice to Sumner's intuitions, at least to some extent.

For perfectly rational individuals wanting is a ‘stronger’ attitude than consenting (at least, in every class of models that satisfies $C-bD$ (Table 3)). If a perfectly rational individual wants A, she also consents to A (given that we accept $C-bD$). However, it is possible for a perfectly rational individual to consent to something that she does not want. A perfectly rational individual may, for example, consent to doing some boring chore in a particular situation even though doing this chore is not something she wants. Sometimes we can use the words ‘agree,’ ‘allow,’ ‘approve,’ ‘condone’ or ‘tolerate’ instead of ‘accept’ or ‘consent.’ Again, in this paper, we use ‘accept’ etc. in an *all-things-considered* sense. It is possible for someone to consent to A even though she objects to some aspects of or consequences of A. In every system in this paper, it is possible (even for a perfectly rational individual) to accept that A and (at the same time) to accept that not-A.

If c is not perfectly rational, almost nothing interesting at all follows logically from the fact that c wants something or accepts something.[5] For instance, if c is not perfectly rational and wants it to be the case that A, it does not follow that it is not the case that c also wants it to be the case that not-A. If c is not perfectly rational and c wants it to be the case that A, it does not follow that c also wants it to be the case that B even if B is a necessary condition for A and c knows this. It is probably not rational to want it to be the case that A and also want it to be the case that not-A, etc., but it does not seem to be *logically* impossible. And, in fact, according to our systems it is not. This is as it should be. We cannot prove the proposition that no one wants it to be the case that A and wants it to be the case that not-A in any system introduced in this paper. Nevertheless, in some systems (for instance every system that includes the rule $T-bD$; Table 14), we can prove the proposition that no one that is *perfectly rational* wants it to be the case that A and wants it to be the case that not-A (for more on this principle, see Section 3.3.3). These facts do not exclude the

[5] I say ‘almost nothing interesting,’ because we can still draw all usual conclusions from this fact. For example, if c wants it to be the case that A, then it is not the case that it is not the case that c wants it to be the case that A, etc. But we do not need a special boulesic logic to draw such conclusions.

possibility that there are *psychological laws* that make it *historically* impossible to combine different attitudes. It might, for instance, be *historically* necessary that no one (not even anyone that is imperfectly rational) wants it to be the case that A and also at the same time wants it to be the case that not-A. I am inclined to believe that this is not the case, at least not always. Yet, it is not *logically* impossible according to our systems: our systems do not rule out this possibility, and I believe that this is plausible.[6]

Boulesic logic, in the sense that I am using the term, is about the *rational* will, not only about what people *actually* want and accept and what follows from this. It attempts to give a description of the will of *perfectly rational*, *reasonable* or *wise* individuals. However, we can also use the systems in this paper to symbolise propositions about what individuals that are not perfectly rational want, accept, etc. Yet, the new, interesting boulesic laws that can be proved in our systems are not empirical, psychological laws; they do not 'describe' the contingent boulesic lives of actual people, even though the contents of these laws can include claims about what individuals that are not perfectly rational want, accept, etc.; they are 'laws of rationality.' Derivatively, boulesic logic also tells us how we must structure our wants if we are to be (perfectly) rational. Of course, exactly what it means to be 'perfectly rational' is something of an open question

[6] It does not appear to be *logically* impossible for human beings to be perfectly rational, but it is very likely that no *actual* human being is perfectly rational; it might even be *historically* impossible for human beings to instantiate this property. Still, it seems to be the case that people are not totally irrational either. Normally, it appears to be true that if someone wants it to be the case that A, she does not also at the same time want it to be the case that not-A, etc. If this were not the case, we would probably not be able to ascribe wants to anyone. Furthermore, it is hard to think that an individual that often wanted it to be the case that A and also wanted it to be the case that not-A would survive for any long period: she would likely be stuck between alternatives like Buridan's ass and starve to death or be eaten by a predator. It is primarily a question for psychologists and other scientists, not for philosophers or logicians, to find out if there are any historically necessary laws of this kind.

and different answers to this question may lead to different boulesic systems. In boulesic logic, we can investigate the consequences of various ways of making the concept of perfect rationality more precise.

There are many ideas about what it means to be rational and many concepts of rationality. According to my view, the 'essence' of rationality is consistency; to be rational is to be consistent (in a wide sense). This involves, at least, consistency with oneself, but perhaps also consistency with the world and with other individuals. According to this view of rationality, it is very plausible to assume that one cannot be perfectly rational if one believes that A and that not-A, or if one wants it to be the case that A and that not-A. This is the core of the concept of rationality that I am trying to explicate in this paper. This concept of rationality should be distinguished from the concept that is often used in, for example, game theory and similar disciplines. In game theory, one usually assumes that every individual is an 'egoist' in the sense that she is only interested in satisfying her own preferences.[7] In game theory rationality is something like enlightened self-interest. Rationality-as-consistency should also be distinguished from a kind of 'rationality' that might be called 'pragmatic.' Suppose an eccentric (and very rich) neuroscientist were to offer you 10.000.000 pounds if you were able to believe in a contradiction and want this contradiction to be true. In this situation it might be plausible to say that it is 'rational' to believe in the contradiction and want it to be true, in some sense of 'rational.' We can call this kind of rationality 'pragmatic.' Suppose you were able to believe in the contradiction and want it to be true. Then we could say that you were pragmatically rational, but you would not be perfectly rational in our sense of this term. Rationality-as-consistency seems to me to be the most basic form, even though I do not deny that it might be fruitful to talk about rationality in other senses too. Much more could be

[7] Such preferences can include otherregarding or altruistic preferences. Still, if an individual does not have any otherregarding or altruistic preferences, it is not rational for her to care about other people according to standard versions of game theory.

said about different theories of rationality, but this suffices for our purposes in the present paper.[8]

One could develop a boulesic logic as a kind of normal modal system and introduce a boulesic operator for every individual, where every operator functions as a normal modal operator. Let us call a boulesic system of this kind an 'ordinary boulesic system.' Nonetheless, there are certain problems with this approach. Firstly, in a system of this kind, it seems unlikely that $\mathcal{W}$ can be used to symbolise what actual persons want, for in such systems everyone wants everything that is necessary, no one has conflicting wants (without wanting absolutely everything), and every individual is such that if she wants A and B is a (logically) necessary means to A (i.e. if A entails B), then she also wants B. But all of this seems false. Secondly, if we restrict a boulesic logic to perfectly rational individuals (to avoid the first problem), we cannot speak about what persons that are not perfectly rational want, accept, etc., at least not in a natural way (we would have to use atomic formulas). Thirdly, we want to be able to symbolise such sentences as 'Everyone wants to be happy,' 'No perfectly rational individual wants it to be the case that both A and not-A,' and 'Everyone in the room wants to take the course.' Fourthly, in an ordinary boulesic system we implicitly have to assume that every perfectly rational individual is necessarily perfectly rational. It is not immediately obvious that this is the case. In our systems, we can investigate what follows if there are individuals that are only contingently perfectly rational. It is also meaningful, in principle, to ask whether we *should* be perfectly rational. If we can only speak about perfectly rational individuals, this does not seem possible. Fifthly, there are many arguments that are intuitively plausible (valid) that cannot be proved in ordinary boulesic systems that can be established in our systems. *Argument* 3 above is an example. The conclusion in this argument is derivable from the premises in every logic in this paper that includes the rule $T-M\mathcal{W}$ (see Table 18). In Section 5.1, I will show this. Yet, *argument* 3 cannot be proved in any ordinary boulesic system, at least not without adding extra, implicit

[8] For more on the concept of rationality, see, for example, (Broome 2013; Horty 2015; and Mele 2004).

premises. For these (and some other) reasons, I think the logics developed in this essay are preferable. In spite of this fact, they can be seen as an elaboration of the modal approach. All our systems include an ordinary modal part, with two kinds of modal operators for absolute and historical necessity and possibility.[9]

As I mentioned in footnote 1, there are some similarities between boulesic logic and 'intentional logic,' even though there are also many important differences. My formal approach is quite different from the formal approach found in the literature on intentional logic. I want to point out some differences. (i) The systems that are developed in the literature on intentional logic are often axiomatic. I use semantic tableaux. (ii) As far as I know, no intentional system includes a distinction between perfectly rational individuals and individuals that are not perfectly rational. Hence, the same logical principles hold for everyone in such systems. In my systems, it is not necessarily the case that individuals who are not perfectly rational satisfy the same principles that perfectly rational individuals satisfy. (iii) Intentions are sometimes required to be consistent while desires are not. In such intentional systems, it is logically impossible that some individual intends to do something and also intends not to do it. In my systems, it is always logically possible that someone (who is not perfectly rational) wants A at the same time that she wants not-A. (iv) Intentional systems are often at least as strong as so called classical modal systems. This means that an individual c intends (that) A iff (if and only if) she intends everything that is logically equivalent with A. In my systems, it is possible that an individual (who is not perfectly rational) wants A even though she does not want everything that is logically equivalent with A. (v) The intentional systems are not usually combined with predicate logic. Therefore, it is not possible to quantify over agents in such systems. They cannot be used to symbolize

[9] For some introductions to ordinary (alethic) modal logic, see, for example, (Blackburn, de Rijke, and Venema 2001; Blackburn, van Benthem, and Wolter 2007; Chellas 1980; Fitting and Mendelsohn 1998; Gabbay 1976; Garson 2006; Kracht 1999; and Lewis and Langford 1932). For more on modal predicate logic, see, for example, (Barcan (Marcus) 1946; Carnap 1946; Garson 1984, 2006; Hintikka 1961; Hughes and Cresswell 1968; Parks 1976; and Priest 2008).

expressions of the following kind: ‘Everyone who is such and such intends to do this or that,’ and ‘someone who is such and such intends to do this or that.’ In all systems in this paper, we can symbolise expressions of the following kind: ‘Everyone who is such and such wants it to be the case that,’ and ‘someone who is such and such wants it to be the case that.’ (vi) Intentional systems are often weaker than so-called normal modal systems (at least for desires). This means that one cannot prove that if an individual desires (intends) (that) A then she also desires (intends) (that) B even though B is a necessary means to A. Similarly, in our systems we cannot prove that if an individual wants A then she also wants every necessary condition of A. However, in some systems we can show that if a *perfectly rational* individual wants A and A necessarily implies B, then she also wants B (see the discussion about hypothetical imperatives below). (vii) At least in some intentional systems the following propositions are valid: if an individual x intends that A and x intends that A implies B then x intends that B, and if it is valid that A implies B then if x intends that A then x intends that B. In our systems, it is not generally true that if an individual x wants it to be the case that A and x wants it to be the case that A implies B then x wants it to be the case that B; nor is it necessarily the case that x wants it to be the case that B if x wants it to be the case that A given that it is valid that A implies B. However, the latter principles do hold in our systems if they are restricted to *perfectly rational* individuals. (viii) As I am using the terms, intentions and wants are not the same thing. You can want someone else to do something, but you cannot intend someone else to do something. Intentions are directed towards (our own) actions, while it is possible to want all sorts of things. Wanting to do something and intending to do it might be the same thing, but it is not obvious that this is the case. If wanting to do something and intending to do it are not the same thing, wanting to do something probably often causes an intention to do it. So, we should make a distinction between intentions and wants. These are some of the most important differences.

Furthermore, I believe, that we should make a distinction between ‘wants’ and ‘wishes’ and not only between ‘wants’ and ‘intentions.’ Wanting something is not (necessarily) the same thing as wishing it were true, even though we may sometimes use ‘wish’ instead of ‘want.’ Wishing something impossible were true might perhaps be possible even for a perfectly rational individual, even though it seems to be reasonable to claim that no perfectly rational individual wants impossible things (in an all-things-considered sense). It seems possible that the sentence ‘I wish you were here’ could be true (even in a situation where it is historically impossible for you to be here [now]), while ‘I want you were here’ is not even grammatical.

There are many good reasons to be interested in the results in this paper, both logical and philosophical. I cannot discuss all of these reasons: instead I will focus on one to illustrate the philosophical usefulness of our technical results.

The systems in this paper can be used to analyse and shed some light upon various interpretations of some philosophically interesting principles—for instance, the so-called *hypothetical imperative.* The notion of a hypothetical imperative was introduced by Immanuel Kant. In *Grundlegung zur Metaphysik der Sitten*, Kant characterises a hypothetical imperative in the following way:

> ‘Who wills the end, wills (so far as reason has decisive influence on his actions) also the means which are indispensably necessary and in his power’ and “‘If I fully will the effect, I also will the action required for it” is analytic.’ (Kant 1991, 45 [originally published in 1785]; English translation in Paton 1948, 80–81.)

Since Kant, there has been debate about how one should formulate the hypothetical imperative and how it should be interpreted, and about whether it is true or not. I will now show how one can use boulesic-deontic logic to distinguish between several different interpretations of this famous principle. I will consider eight of the most interesting readings and then show how they can be formalised in our systems. Finally, I will indicate which versions can be proved in various systems.[10]

1. It is universally necessary that, for every x, if x wants it to be the case that A and it is necessary that if A then B, then x wants it to be the

[10] Kant is usually taken to mean that moral principles are necessary and universal. Hence, I will interpret the hypothetical imperative in the same way.

case that B. $U\Pi x((\mathcal{W}_x A \land \Box(A \rightarrow B)) \rightarrow \mathcal{W}_x B)$. (Translation key. U: It is universally (absolutely) necessary that. Πx: For every [possible] x. $\mathcal{W}_x$: x wants it to be the case that. $\Box$: It is historically necessary that. $\rightarrow$: Material implication.)

2. It is universally necessary that, for every x, if x wants it to be the case that A and it is necessary that if A then B, then it ought to be the case that B. $U\Pi x((\mathcal{W}_x A \land \Box(A \rightarrow B)) \rightarrow OB)$. (Translation key. The same as before. O: It ought to be the case that.)

3. It is universally necessary that, for every x, if x wants it to be the case that A and it is necessary that if A then B, then x ought to want it to be the case that B. $U\Pi x((\mathcal{W}_x A \land \Box(A \rightarrow B)) \rightarrow O\mathcal{W}_x B)$.

4. It is universally necessary that, for every x, it ought to be the case that if x wants it to be the case that A and it is necessary that if A then B, then x wants it to be the case that B. $U\Pi x O((\mathcal{W}_x A \land \Box(A \rightarrow B)) \rightarrow \mathcal{W}_x B)$.

5. It is universally necessary that, for every x, it ought to be the case that if x wants it to be the case that A and it is necessary that if A then B, then B. $U\Pi x O((\mathcal{W}_x A \land \Box(A \rightarrow B)) \rightarrow B)$.

6. It is universally necessary that, for every x, if x is perfectly rational, then if x wants it to be the case that A and it is necessary that if A then B, then x wants it to be the case that B. $U\Pi x(Rx \rightarrow ((\mathcal{W}_x A \land \Box(A \rightarrow B)) \rightarrow \mathcal{W}_x B))$.

7. It is universally necessary that, for every x, if x is perfectly rational, then if x wants it to be the case that A and it is necessary that if A then B, then it ought to be the case that B. $U\Pi x(Rx \rightarrow ((\mathcal{W}_x A \land \Box(A \rightarrow B)) \rightarrow OB))$.

8. It is universally necessary that, for every x, if x is perfectly rational, then if x wants it to be the case that A and it is necessary that if A then B, then x ought to want it to be the case that B. $U\Pi x(Rx \rightarrow ((\mathcal{W}_x A \land \Box(A \rightarrow B)) \rightarrow O\mathcal{W}_x B))$.

(1), (2), (3), (6), (7) and (8) are so-called 'narrow-scope' readings of the hypothetical imperative; (4) and (5) are so-called 'wide-scope' readings. Note how the consequent in the various interpretations varies. In (1), the consequent is about an attitude; in (2), the consequent is about a norm; and in (3), it is about a norm about an attitude, etc. Readings (1)–(5) cannot be proved in any system in this paper: they are not valid in any class of model that we consider. In Section 5.1, I will show that the following 'instance' of (1) is not valid (in the class of all models): $\Pi x((\mathcal{W}_x Qx \land \Box(Qx \to Dx)) \to \mathcal{W}_x Dx)$, where Q and D are two monadic predicates. In our example, Qx stands for 'x quenches her thirst' and Dx stands for 'x drinks some water.' Since $\Pi x((\mathcal{W}_x Qx \land \Box(Qx \to Dx)) \to \mathcal{W}_x Dx)$ is not valid in the class of all models, it follows that not every instance of $U\Pi x((\mathcal{W}_x A \land \Box(A \to B)) \to \mathcal{W}_x B)$ is valid in the class of all models. Nevertheless, (6)–(8) can be deduced in some systems: (6) is provable in any boulesic (or boulesic-deontic) system that includes the rule $T-M\mathcal{W}$; (7) and (8) cannot be established in any pure boulesic system: we need a boulesic-deontic system; (7) can be derived in every boulesic-deontic system that includes the rules $T-\mathcal{W}O$, $T-H\mathcal{W}$ and $T-M\mathcal{W}$ (I will establish this in Section 5.1); (8) can be proved in every boulesic-deontic system that includes the rules $T-\mathcal{W}O$, $T-H\mathcal{W}$, $T-M\mathcal{W}$, $T-a4$, $T-b4$ and $T-FTR$. (For more on these rules, see Section 4.2.)

It seems to me that interpretation (6) comes very close to Kant's own reading of the hypothetical imperative. If this is correct, we can prove that Kant was right: 'Who wills the end, wills (so far as reason has decisive influence on his actions) also the means which are indispensably necessary...' is, in this sense, a necessary universal truth in some models. In Section 3.6, I will discuss a semantic argument that shows that (6) is valid in the class of all models that satisfy $C-M\mathcal{W}$ (see Table 5).[11]

[11] For more information about the hypothetical imperative, see, for example, (Bedke 2009; Broome 1999; Brunero 2010; Downie 1984; Feldman 1986, Chapter 5; Foot 1972; Gensler 1985; Greenspan 1975; Harsanyi 1958; Hill 1973, 1989; Korsgaard 2008; Marshall 1982; Shaver 2006; Schroeder 2004, 2005, 2009, 2015; Wallace 2001; and Way 2010). There are at least two interesting questions about the interpretation of the hypothetical imperative. (1) What does (or should) the 'consequent' of the hypothetical imperative say (is it a claim about an attitude, a norm about what ought to be or about what we ought to do, or a norm about an attitude)? Kant seems to think that the consequent is about an attitude, about willing. But it has also been suggested that the consequent is a norm about what we ought to do or

The discussion of the hypothetical imperative above clearly shows, I believe, that the systems introduced in this paper are philosophically interesting. In conclusion, the topic of this article is both philosophically and logically well motivated.[12]

a norm about an attitude. Hill (1973), for instance, suggests that the consequent might be about what 'we' *ought to want*, and Marshall (1982) and Schroeder (2004) suggest that it might be about what 'we' *ought to do*. (2) Should the imperative be given a wide-scope or a narrow-scope interpretation? There is a debate about what Kant meant. Many philosophers seem to prefer a wide-scope reading of hypothetical imperatives, see e.g. (Hill 1973; Gensler 1985; Wallace 2001; Broome 1999, 2001; Greenspan 1975). But some have also argued for a narrow-scope reading, see e.g. (Schroeder 2004, 2005, 2009). According to Schroeder, Kant should be interpreted as a narrow-scoper. I am inclined to believe that this is in fact a better interpretation of Kant's position. Whether or not this view is correct, it is a nice feature of the systems in this paper that we can clearly distinguish between these different readings.

[12] There are many other good reasons to be interested in the systems in this paper and also some potential problems with the whole project. I cannot discuss every interesting philosophical issue that is related to the topics of this paper. However, I would like to briefly mention a potential problem that was raised by an anonymous reviewer. According to this reviewer it may in principle be interesting to devote some attention to specific logical/inferential features of sentences which speak about willing/wanting, but there is no real need for a comprehensive logical theory of expressions of this kind. It is, of course, possible that this view is correct, but is seems highly problematic to me. If it is interesting to devote some attention to specific logical/inferential features of sentences which speak about willing/wanting, it seems to me that it must also be interesting to try to develop a comprehensive logical theory of expressions of this kind. In general, it is more interesting to have a comprehensive theory of some 'phenomenon' than just a piecemeal grasp of some unconnected truths (and almost everyone, I think, agrees with this view). Consider a similar argument. 'Though in principle it may be interesting to devote some attention to specific logical/inferential features of sentences which speak about propositional (truth-functional) truths there is no real need for a comprehensive logical theory of expressions of this kind. For example, we do not need sound and complete systems of propositional logic. There is no point in trying to construct axiomatic systems or tableau systems of propositional logic. It is enough if we study the law of non-contradiction, the law of identity, the law of excluded middle, etc.' This argument is obviously highly problematic and I believe few people would be convinced by it. When philosophers and logicians started to study propositional logic systematically, constructed sound and complete systems, and proved that they were sound and

The paper is divided into seven main sections. Section 2 deals with the syntax and Section 3 with the semantics of our systems. In Section 4, I describe the proof theory of our logics, while Section 5 includes some examples of theorems. Section 6 contains soundness and completeness proofs for every system. Finally, Section 7 includes a short conclusion and summary.

2. Syntax

2.1. Alphabet

Terms

(i) A set of variables x_1, x_2, x_3, . . .

(ii) A set of constants (rigid designators) k_{d_1}, k_{d_2}, k_{d_3}, . . .

Predicates

(iii) For every natural number $n > 0$, n-place predicate symbols P^1_n, P^2_n, P^3_n, . . .

(iv) The monadic existence predicate E, and the monadic rationality predicate R.

(v) The dyadic identity predicate (necessary identity) =.

Connectives

(vi) The primitive truth-functional connectives $\neg$ (negation), $\wedge$ (conjunction), $\vee$ (disjunction), $\rightarrow$ (material implication) and $\leftrightarrow$ (material equivalence).

complete, this was an extremely important development in the history of logic. Why should there be any difference if one talks about ‘logical/inferential features of sentences which speak about willing/wanting’? The reviewer might be right that we do not ‘need’ boulesic logic in some senses of this word. For example, we do not need it to survive or for society to go on functioning. But, then again, we do not need *any* comprehensive logical theory at all for these purposes. The fact that we do not ‘need’ boulesic logic in some senses of the term ‘need,’ doesn’t show that the topic of my paper isn’t interesting. And it certainly does not follow that the project is not worth the effort. In conclusion, this potential problem does not strike me as particularly serious.

Operators

(vii) The alethic operators U (universal necessity), M (universal possibility), $\Box$ (historical necessity) and $\Diamond$ (historical possibility).[13]

(viii) The deontic operators O (Ought) and P (Permission).

(ix) The boulesic operators $\mathcal{W}$ (Want) and $\mathcal{A}$ (Accept).

Quantifiers

(x) The (possibilist) quantifiers Π (For all) and Σ (For some).

Parentheses

(xi) The brackets) and (.

I will use x, y and $z \ldots$ for arbitrary variables, a, b, $c \ldots$ for arbitrary constants, and s and t for arbitrary terms (with or without primes or subscripts). For more on the set of constants, see Section 3.1. I will use F_n, G_n, $H_n, \ldots$ for arbitrary n-place predicates and I will omit the subscript if it can be read off from the context.

Π and Σ are substitutional, 'possibilist' quantifiers because the domain is the same in every possible world and every object in the domain has a name (Section 3). Thus, in effect, they vary over every object in the domain.

2.2. Languages

I will consider two languages in this paper. The first, $L1$, does not include the deontic operators; the second, $L2$, does. $L1$ is constructed from clauses (i)–(viii) and (x), and $L2$ from clauses (i)–(x) below.

(i) Any constant or variable is a term.

(ii) If $t_1, \ldots, t_n$ are any terms and P is any n-place predicate, $Pt_1 \ldots t_n$ is an atomic formula.

[13] U and M are standard universal modalities (see almost any introduction to modal logic). For more on the concepts of historical necessity and possibility, see, for example, (Åqvist and Hoepelman 1981; and Chellas 1969). In this paper, I will not try to combine boulesic logic and temporal logic since I want to keep things relatively simple. However, it is in principle possible to combine the systems in this paper with various tense systems. I hope to do this in future work (see the conclusion in Section 7).

(iii) If t is a term, Et (‘t exists’) is an atomic formula and Rt (‘t is perfectly rational’) is an atomic formula.

(iv) If s and t are terms, then $s = t$ (‘s is identical with t’) is an atomic formula.

(v) If A and B are formulas, so are $\neg A$, $(A \wedge B)$, $(A \vee B)$, $(A \rightarrow B)$ and $(A \leftrightarrow B)$.

(vi) If A is a formula, so are UA (‘it is universally [or absolutely] necessary that A’), MA (‘it is universally [or absolutely] possible that A’), $\Box A$ (‘it is [historically] necessary that A’) and $\Diamond A$ (‘it is [historically] possible that A’).

(vii) If B is any formula and t is any term, then $\mathcal{W}_t B$ (‘t wants it to be the case that B’) and $\mathcal{A}_t B$ (‘t accepts that it is the case that B’) are formulas.

(viii) If A is any formula and x is any variable, then $\Pi x A$ (‘for every [possible] x: A’) and $\Sigma x A$ (‘for some [possible] x: A’) are formulas.

(ix) If A is a formula, then OA (‘it ought to be the case that A’) and PA (‘it is permitted that A’) are formulas.

(x) Nothing else is a formula.

A, B, C ... stand for arbitrary formulas, and Γ, Φ ... for sets of formulas. The concepts of bound and free variables, and open and closed formulas, are defined in the usual way. $(A)[t/x]$ is the formula obtained by substituting t for every free occurrence of x in A. The definition is standard. Brackets around formulas are usually dropped if the result is not ambiguous.

2.3. Definitions

It is possible to introduce some new symbols into our languages by definitions. If we do that, the new symbols should be treated as pure metalogical abbreviations and we should not read anything more into the definitions. Here are some examples:

Deontic operators. $FA =df\, \neg PA$. $KA =df\, (PA \wedge P\neg A)$. $NA =df\, \neg KA$.

Actualist quantifiers. $\forall x A =df\, \Pi x(Ex \rightarrow A)$ and $\exists x A =df\, \Sigma x(Ex \wedge A)$.

O and P are not included in $L1$; in $L2$ they are treated as primitive symbols. In some (but not all) systems that I will describe, O, P and F are 'definable' in terms of the boulesic operators in the sense that we can prove that the following equivalences are logically true: $OB \leftrightarrow \Pi x(Rx \rightarrow \mathcal{W}_x B)$ ('It ought to be the case that B iff everyone who is perfectly rational wants it to be the case that B'); $PB \leftrightarrow \Pi x(Rx \rightarrow \mathcal{A}_x B)$ ('It is permitted that B iff everyone who is perfectly rational accepts that [consents to the state of affairs that] it is the case that B'); and $FB \leftrightarrow \Pi x(Rx \rightarrow \mathcal{W}_x \neg B)$ ('It is not permitted that B iff everyone who is perfectly rational wants it to be the case that not-B'). However, since it is not immediately obvious that these equivalences *should* hold, and since we want to know which assumptions we must make to be able to prove them, we do not introduce the deontic operators through definitions in this paper. Furthermore, when I say that O, P and F are 'definable' in terms of the boulesic operators, this should not be taken to imply that, for example, 'OB' has the same meaning as '$\Pi x(Rx \rightarrow \mathcal{W}_x B)$' or that '$OB$' can be replaced by '$\Pi x(Rx \rightarrow \mathcal{W}_x B)$' in every context. (For more on this, see Section 4.2.12 and Table 30.)[14]

[14] One possible objection against these equivalences is that they seem to presuppose an extreme view of rationality according to which all rational agents should have essentially the same wishes. It is true that in the systems where we can prove the equivalences, all perfectly rational individuals want the same things (see Sections 3.3.7, 4.2.8 and 4.2.12). If a is perfectly rational and b is perfectly rational, then it is not the case that a wants C and b wants not-C in those systems (at least, if we also assume that they include, for example, $T - bD$ (Table 14)). But is it not the case that this leaves no scope for legitimate conflicts of interests such as two businesspeople each wishing their own company to take market shares from that of the other, or two fathers both wishing that their own child wins a competition? I cannot discuss this argument in detail in this paper, but I want to make the following remarks. (1) Even in systems where we can prove the equivalences, it is possible that an individual a wants C and another individual b wants not-C (given that not both a and b are perfectly rational). (2) Even in systems where we can prove the equivalences, it is possible that it is permitted (and even obligatory) for some individual a to want C and for some other individual b to want not-C. (3) Even in systems where we can prove the equivalences and where both a and b are perfectly rational, it is possible that a wants to do 'everything' a can to win and that b wants to do 'everything' b can to win and that both a and b want 'the best man' to win, even

3. Semantics

3.1. Models

Definition 1 *A model $\mathcal{M}$ is a relational structure $\langle D, W, \mathfrak{R}, \mathfrak{A}, v\rangle$, where D is a non-empty set of individuals (the domain), W is a non-empty set of possible worlds, $\mathfrak{R}$ is a binary alethic accessibility relation ($\mathfrak{R}$ is a subset of $W \times W$), $\mathfrak{A}$ is a ternary boulesic accessibility relation ($\mathfrak{A}$ is a subset of $D \times W \times W$), and v is an interpretation function.*

A supplemented model $\mathcal{M}_S$ is a relational structure $\langle D, W, \mathfrak{R}, \mathfrak{S}, \mathfrak{A}, v\rangle$, where D, W, $\mathfrak{R}$, $\mathfrak{A}$ and v are as in an ordinary model, and $\mathfrak{S}$ is a dyadic deontic accessibility relation ($\mathfrak{S}$ is a subset of $W \times W$).

$\mathfrak{R}$ 'corresponds' to the alethic operators $\Box$ and $\Diamond$, $\mathfrak{S}$ to the deontic operators O and P, and $\mathfrak{A}$ to the boulesic operators $\mathcal{W}$ and $\mathcal{A}$. Informally, $\mathfrak{R}\omega\omega'$ says that the possible world ω' is alethically (historically) accessible from the possible world ω, $\mathfrak{S}\omega\omega'$ says that the possible world ω' is deontically accessible from the possible world ω, and $\mathfrak{A}\delta\omega\omega'$ says that the possible world ω' is acceptable to the individual δ in (or relative to) the possible world ω, or that δ accepts ω' in (or relative to) ω.

In Section 3.4, we will see how $\mathfrak{S}$ and $\mathfrak{A}$ can be defined, and we will explore the consequences of these definitions.

The valuation function v assigns every constant c an element $v(c)$ of D, and every possible world ω in W and an n-place predicate P a subset $v_\omega(P)$ (the extension of P in ω) of D^n. In other words, $v_\omega(P)$ is the set of n-tuples that satisfy P in the world ω. Hence, every constant is a kind of rigid designator: it refers to the same individual in every possible world. Nonetheless, the extension of a predicate may change from world to world and it may be empty in a world. Let $\mathcal{M}$ be an ordinary or supplemented model. Then the

though it is not possible that both a and b win. (4) All systems I discuss are compatible with the proposition that it is possible for perfectly rational individuals to *wish* for incompatible things even though it is not possible for them to *want* incompatible things. So, it is not the case that those systems leave *no* scope for legitimate conflicts of interests. Whether or not they leave *enough* scope is, of course, debatable. Personally, I am inclined to believe that they *do* leave enough scope.

language of $\mathcal{M}$, $\mathcal{L}(\mathcal{M})$, is obtained by adding a constant k_d, such that $v(k_d) = d$, to the language for every member $d \in D$. Hence, every object in the domain of a model has at least one name in our language, but several different constants may refer to one and the same object.

The predicate R has a special interpretation in our systems. 'Rc' says that c is *perfectly rational*, *perfectly reasonable* or *perfectly wise*. If $v(c)$ is in the extension of R at the possible world ω, this means that $v(c)$ is perfectly rational, reasonable or wise in ω. Exactly what this means will depend on the conditions we impose on the boulesic accessibility relation $\mathfrak{A}$ (Section 3.3). R functions as an ordinary predicate. Hence, an individual δ may be in R's extension in one possible world even though δ is not in R's extension in every possible world. Accordingly, the fact that an individual δ is perfectly rational, reasonable or wise in a possible world does not *entail* that δ is perfectly rational, reasonable or wise in *every* possible world. In Section 3.3.8, we will see what happens if we add the extra assumption that every perfectly rational individual is necessarily perfectly rational (the semantic condition *C–UR* guarantees that this is the case: see Table 8). In the light of the definitions of the truth conditions for sentences of the forms $\mathcal{W}_a C$ and $\mathcal{A}_a C$ (see Section 3.2, conditions (ii), (xi) and (xii)), it should be obvious that R plays an important role in our systems. It will become even clearer when we introduce the various tableau rules in Section 4.2. Whether or not we can draw any interesting consequences from the fact that an individual c wants (or accepts) something (in a possible world) will depend on whether or not c is perfectly rational, that is, whether or not c is in R's extension (in this possible world).

The valuation function assigns extensions to so-called matrices. Given any closed boulesic formula of the form $\mathcal{W}_t C$ or $\mathcal{A}_t C$, we shall construct its matrix as follows. Let m be the least number greater than every n such that x_n occurs bound in C. From left to right, replace every occurrence of an individual constant with x_m, x_{m+1}, etc. The result is the formula's matrix. Here are some examples: the matrix of $\mathcal{W}_d Pc$ is $\mathcal{W}_{x_1} Px_2$; the matrix of $\mathcal{A}_c Pdc$ is $\mathcal{A}_{x_1} Px_2x_3$; the matrix of $\mathcal{W}_c(Pa \leftrightarrow (Pa \wedge Pa))$ is $\mathcal{W}_{x_1}(Px_2 \leftrightarrow (Px_3 \wedge Px_4))$; the matrix of $\mathcal{W}_c \Sigma_{x_1}(Fc \rightarrow G_{x_1})$ is $\mathcal{W}_{x_2} \Sigma x_1(Fx_3 \rightarrow Gx_1)$; the matrix of $\mathcal{W}_c\mathcal{W}_d\Pi x_3Px_3$ is $\mathcal{W}_{x_4}\mathcal{W}_{x_5}\Pi x_3Px_3$, the matrix of $\mathcal{W}_c\Pi x_1\mathcal{W}_{x_1}\Sigma x_2Px_1x_2$ is $\mathcal{W}_{x_3}\Pi x_1\mathcal{W}_{x_1}\Sigma x_2Px_1x_2$, etc.

Let A be any formula. Then, $(A)[a_1, \ldots, a_n/x_1, \ldots, x_n]$ is the result of replacing every free occurrence of x_1 by a_1, and ..., and every free occurrence of x_n by a_n in A. $(A)[a_1, \ldots, a_n/x_1, \ldots, x_n]$ will be abbreviated as $(A)[a_1, \ldots, a_n/\vec{x}]$ (parentheses around A will sometimes be dropped). Here are some examples. Let A be $\mathcal{W}_{x_1}Px_2$. Then, $(A)[d,\ c/x_1,\ x_2] = \mathcal{W}_dPc$. Let A be $\mathcal{A}_{x_1}Px_2x_3$. Then, $(A)[c,\ d,\ c/x_1,\ x_2,\ x_3] = \mathcal{A}_cPdc$. Let A be $\mathcal{W}_{x_4}\mathcal{W}_{x_5}\Pi x_3Px_3$. Then, $(A)[c,\ d/x_4,\ x_5] = \mathcal{W}_c\mathcal{W}_d\Pi x_3Px_3$, etc.

If M is any matrix of the form $\mathcal{W}_tC$ or $\mathcal{A}_tC$ with free variables $x_1, \ldots, x_n$, then $v_\omega(M) \subseteq D^n$. Intuitively, this means that M is interpreted as a predicate and not as a (closed) sentence. Note that M always includes at least one free variable. Let M be a matrix where x_m is the first free variable in M and a_m is the constant in $M[a_1, \ldots, a_n/\vec{x}]$ that replaces x_m. Then the truth conditions for closed boulesic formulas of the form $M[a_1, \ldots, a_n/\vec{x}]$, when $v_\omega(Ra_m) = 0$, are defined in terms of the extension of M in ω. If $v_\omega(Rc) = 1$, then $\mathcal{W}_c$ in $\mathcal{W}_cB$ ($\mathcal{A}_c$ in $\mathcal{A}_cB$) will behave as a modal operator in ω. (See conditions (ii), (xi) and (xii) in Section 3.2 below for more on this.)[15]

$v_\omega(=) = \{\langle d,\ d\rangle : d \in D\}$, i.e. the extension of the identity predicate is the same in every possible world (in a model). It follows that all identities (and non-identities) are both absolutely and historically necessary. The existence predicate E functions as an ordinary predicate. The extension of this predicate may vary from one world to another. 'Ec' is true in a possible world iff $v(c)$ exists in this world.

3.2. Truth conditions

We now extend the interpretation function. Every closed formula, A, is assigned exactly one truth-value (1 = True or 0 = False), $v_\omega(A)$, in each world ω.

Here are the truth conditions for some sentences in our language. (The truth conditions for the omitted truth-functional connectives are the usual

[15] See Priest (2005, Ch. 1–2) and Section 5.1 in this paper for more on matrices.

ones. '$\forall\omega' \in W$' means 'for every possible world ω' in W'; and '$\exists\omega' \in W$' means 'for some possible world ω' in W.')

(i) $v_\omega(Pa_1 \dots a_n) = 1$ iff $\langle v(a_1), \dots, v(a_n)\rangle \in v_\omega(P)$.

Let M be a matrix where x_m is the first free variable in M and a_m is the constant in $M[a_1, \dots, a_n/\vec{x}]$ that replaces x_m. Then the truth conditions for closed boulesic formulas of the form $M[a_1, \dots, a_n/\vec{x}]$, when $v_\omega(Ra_m) = 0$, are given in (ii) below.

(ii) $v_\omega(M[a_1, \dots, a_n/\vec{x}]) = 1$ iff $\langle v(a_1), \dots, v(a_n)\rangle \in v_\omega(M)$.
(iii) $v_\omega(UA) = 1$ iff $\forall\omega' \in W$: $v_{\omega'}(A) = 1$.
(iv) $v_\omega(MA) = 1$ iff $\exists\omega' \in W$: $v_{\omega'}(A) = 1$.
(v) $v_\omega(\Box A) = 1$ iff $\forall\omega' \in W$ s.t. $\mathfrak{R}\omega\omega'$: $v_{\omega'}(A) = 1$.
(vi) $v_\omega(\Diamond A) = 1$ iff $\exists\omega' \in W$ s.t. $\mathfrak{R}\omega\omega'$: $v_{\omega'}(A) = 1$.
(vii) $v_\omega(OA) = 1$ iff $\forall\omega' \in W$ s.t. $\mathfrak{S}\omega\omega'$: $v_{\omega'}(A) = 1$.
(viii) $v_\omega(PA) = 1$ iff $\exists\omega' \in W$ s.t. $\mathfrak{S}\omega\omega'$: $v_{\omega'}(A) = 1$.
(ix) $v_\omega(\Pi xA) = 1$ iff for all $k_d \in \mathcal{L}(\mathcal{M})$, $v_\omega(A[k_d/x]) = 1$.
(x) $v_\omega(\Sigma xA) = 1$ iff for some $k_d \in \mathcal{L}(\mathcal{M})$, $v_\omega(A[k_d/x]) = 1$.

Note that O and P are not included in the language $L1$, while they are primitive in the language $L2$.

Here are the truth conditions for $\mathcal{W}_aC$ and $\mathcal{A}_aC$.

(xi) $v_\omega(\mathcal{W}_aC) = 1$ iff for all ω' such that $\mathfrak{A}v(a)\omega\omega'$: $v_{\omega'}(C) = 1$, given that $v(a)$ is an element in $v_\omega(R)$, if $v(a)$ is not an element in $v_\omega(R)$, then $\mathcal{W}_aC$ is assigned a truth value in ω in a way that does not depend on the value of C (see condition (ii) above).
(xii) $v_\omega(\mathcal{A}_aC) = 1$ iff for at least one ω' such that $\mathfrak{A}v(a)\omega\omega'$: $v_{\omega'}(C) = 1$, given that $v(a)$ is an element in $v_\omega(R)$, if $v(a)$ is not an element in $v_\omega(R)$, then $\mathcal{A}_aC$ is assigned a truth value in ω in a way that does not depend on the value of C (see condition (ii) above).

Intuitively, conditions (xi) and (xii) can be interpreted in the following way: if $v(a)$ is not perfectly rational in a possible world, $\mathcal{W}_aC$ and $\mathcal{A}_aC$ behave as if they are ordinary predicates in this world; and if $v(a)$ is perfectly rational in a possible world, $\mathcal{W}_a$ and $\mathcal{A}_a$ behave as ordinary modal operators in this world. So, the truth value of $\mathcal{W}_aC$ ($\mathcal{A}_aC$) in the possible world ω when a is not perfectly rational in ω is not determined by anything that goes on in some other world. It is, for example, logically possible for someone who is not perfectly rational to want C without wanting B even though B is a necessary means to C (see Section 5.1 for more on this).

3.3. Conditions on models

In this section, I will consider some conditions that might be imposed on our models. These conditions concern the formal properties of the accessibility relations, the relationships between the various accessibility relations and the relationships between the accessibility relations and the valuation function. In the formulas in this section, we can think of c and d as varying over individuals in D, and x, y, z and w as varying over possible worlds in W. Table 1 and Table 2 include information about the alethic and the deontic accessibility relations. The well-known conditions introduced in these tables are mentioned in almost any introduction to modal and deontic logic (see the introduction for some references). The clauses in Table 4, which concern the relationships between the alethic and the deontic accessibility relations, have been discussed by Rönnedal (2012), for instance. All other conditions are new.

The clauses in this section can be combined in many different ways, generating many different boulesic and boulesic-deontic systems (sections 3.5 and 4.3). Exactly which conditions *should* we accept? The answer to this question will depend on what it means to be perfectly rational. I think there might be good reasons to accept *all* (or almost all) conditions in this section. In Section 3.4, I will consider one such reason. However, it might also be interesting to see what follows if we accept some other, smaller class. The conditions in this section should be more or less self-explanatory. Nevertheless, I have added a few comments about some of the new clauses. There are many interesting relationships between the various conditions that I do not have space to discuss in this paper. In Section 3.4, I will consider what follows if we define the ternary boulesic accessibility relation $\mathfrak{A}$ in terms of the alethic accessibility relation $\mathfrak{R}$ and a binary acceptance predicate, and the deontic accessibility relation $\mathfrak{S}$ in terms of the ternary boulesic accessibility relation. This will show how deontic logic can in a certain sense be grounded in boulesic logic.

3.3.1. Conditions on the relation $\mathfrak{R}$

Condition	Formalisation of condition
$C-aT$	$\forall x \mathfrak{R}xx$
$C-aD$	$\forall x \exists y \mathfrak{R}xy$
$C-aB$	$\forall x \forall y(\mathfrak{R}xy \rightarrow \mathfrak{R}yx)$
$C-a4$	$\forall x \forall y \forall z((\mathfrak{R}xy \wedge \mathfrak{R}yz) \rightarrow \mathfrak{R}xz)$
$C-a5$	$\forall x \forall y \forall z((\mathfrak{R}xy \wedge \mathfrak{R}xz) \rightarrow \mathfrak{R}yz)$

Table 1

3.3.2. Conditions on the relation $\mathfrak{S}$

Condition	Formalisation of condition
$C-dD$	$\forall x \exists y \mathfrak{S}xy$
$C-d4$	$\forall x \forall y \forall z((\mathfrak{S}xy \wedge \mathfrak{S}yz) \rightarrow \mathfrak{S}xz)$
$C-d5$	$\forall x \forall y \forall z((\mathfrak{S}xy \wedge \mathfrak{S}xz) \rightarrow \mathfrak{S}yz)$
$C-dT'$	$\forall x \forall y(\mathfrak{S}xy \rightarrow \mathfrak{S}yy)$
$C-dB'$	$\forall x \forall y \forall z((\mathfrak{S}xy \wedge \mathfrak{S}yz) \rightarrow \mathfrak{S}zy)$

Table 2

3.3.3. Conditions on the relation $\mathfrak{A}$

Condition	Formalisation of condition
$C-bD$	$\forall d \forall x \exists y \mathfrak{A}dxy$
$C-b4$	$\forall d \forall x \forall y \forall z((\mathfrak{A}dxy \wedge \mathfrak{A}dyz) \rightarrow \mathfrak{A}dxz)$
$C-b5$	$\forall d \forall x \forall y \forall z((\mathfrak{A}dxy \wedge \mathfrak{A}dxz) \rightarrow \mathfrak{A}dyz)$
$C-bT'$	$\forall d \forall x \forall y(\mathfrak{A}dxy \rightarrow \mathfrak{A}dyy)$
$C-bB'$	$\forall d \forall x \forall y \forall z((\mathfrak{A}dxy \wedge \mathfrak{A}dyz) \rightarrow \mathfrak{A}dzy)$

Table 3[16]

[16] 'C' in '$C-bD$' stands for 'condition' and 'b' for 'boulesic.' $C-bD$ is called '$C-bD$' because it is similar to the well-known condition D (as in 'deontic') in ordinary

The conditions in Table 3 correspond to the tableau rules in Table 14. Note that $\mathfrak{A}$ is a ternary relation. $C\!-bD$ (Table 3) says: for every (individual) d and for every (possible world) x there is a (possible world) y such that d accepts y in x. According to this condition, every individual always accepts at least one possible world, no matter what situation she is in. If all possibilities are in some sense 'bad,' she will accept the possibility (or possibilities) that is (are) 'least bad,' so to speak. In all classes of models that satisfy this condition, the following sentence (schema) is valid: $\Pi x(Rx \to \neg(\mathcal{W}_x B \wedge \mathcal{W}_x \neg B))$ ('For every x: if x is perfectly rational, then it is not the case that x wants it to be the case that B and x wants it to be the case that not-B'). This is an intuitively plausible principle. If c wants it to be the case that B and also wants it to be the case that not-B, not all of c's wants can be satisfied. There is no possible world in which both B and not-B are true, and c cannot see to it that B and see to it that not-B.

3.3.4. Conditions concerning the relation between $\mathfrak{R}$ and $\mathfrak{S}$

Condition	Formalisation of condition
$C\!-MO$	$\forall x \forall y(\mathfrak{S}xy \to \mathfrak{R}xy)$
$C\!-OC$	$\forall x \exists y(\mathfrak{S}xy \wedge \mathfrak{R}xy)$
$C\!-OC'$	$\forall x \forall y(\mathfrak{S}xy \to \exists z(\mathfrak{S}yz \wedge \mathfrak{R}yz))$
$C\!-MO'$	$\forall x \forall y \forall z((\mathfrak{S}xy \wedge \mathfrak{S}yz) \to \mathfrak{R}yz)$
$C\!-ad4$	$\forall x \forall y \forall z((\mathfrak{R}xy \wedge \mathfrak{S}yz) \to \mathfrak{S}xz)$
$C\!-ad5$	$\forall x \forall y \forall z((\mathfrak{R}xy \wedge \mathfrak{S}xz) \to \mathfrak{S}yz)$
$C\!-PMP$	$\forall x \forall y \forall z((\mathfrak{S}xy \wedge \mathfrak{R}xz) \to \exists w(\mathfrak{R}yw \wedge \mathfrak{S}zw))$
$C\!-OMP$	$\forall x \forall y \forall z((\mathfrak{R}xy \wedge \mathfrak{S}yz) \to \exists w(\mathfrak{S}xw \wedge \mathfrak{R}wz))$

Table 4

alethic (modal) logic. Similar remarks apply to the other conditions in this section. It is usually binary relations that are called serial, transitive, Euclidean, etc. Nonetheless, we will extend these concepts to ternary relations. If $\mathfrak{A}$ satisfies $C\!-b4$, we will call $\mathfrak{A}$ transitive, and so on. If it is clear from the context that we are talking about a semantic condition, I will often omit the initial C.

3.3.5. Conditions concerning the relation between $\mathfrak{R}$ and $\mathfrak{A}$

Condition	Formalisation of condition
$C- M\mathcal{W}$	$\forall d \forall x \forall y(\mathfrak{A}dxy \to \mathfrak{R}xy)$
$C- \mathcal{W}C$	$\forall d \forall x \exists y(\mathfrak{A}dxy \wedge \mathfrak{R}xy)$
$C- \mathcal{W}C'$	$\forall d \forall x \forall y(\mathfrak{A}dxy \to \exists z(\mathfrak{A}dyz \wedge \mathfrak{R}yz))$
$C- M\mathcal{W}'$	$\forall d \forall x \forall y \forall z((\mathfrak{A}dxy \wedge \mathfrak{A}dyz) \to \mathfrak{R}yz)$
$C- ab4$	$\forall d \forall x \forall y \forall z((\mathfrak{R}xy \wedge \mathfrak{A}dyz) \to \mathfrak{A}dxz)$
$C- ab5$	$\forall d \forall x \forall y \forall z((\mathfrak{R}xy \wedge \mathfrak{A}dxz) \to \mathfrak{A}dyz)$
$C- \mathcal{A}MP$	$\forall d \forall x \forall y \forall z((\mathfrak{A}dxy \wedge \mathfrak{R}xz) \to \exists w(\mathfrak{R}yw \wedge \mathfrak{A}dzw))$
$C- \mathcal{W}MP$	$\forall d \forall x \forall y \forall z((\mathfrak{R}xy \wedge \mathfrak{A}dyz) \to \exists w(\mathfrak{A}dxw \wedge \mathfrak{R}wz))$

Table 5[17]

The conditions in Table 5 are similar to the conditions in Table 4. However, the clauses in Table 5 concern the relationship between the boulesic accessibility relation and the alethic accessibility relation. The conditions in Table 5 correspond to the tableau rules in Table 18. $C- M\mathcal{W}$ says: 'For every (individual) d, for every (possible world) x and for every (possible world) y, d accepts y in x only if y is alethically accessible from x.' In other words, if $C- M\mathcal{W}$ holds, then it is not the case that d accepts y in x if y is not alethically accessible from x. In every class of models that satisfies this condition, the following version of the hypothetical imperative is valid: $U\Pi x(Rx \to ((\mathcal{W}_x A \wedge \Box(A \to B)) \to \mathcal{W}_x B))$ (see the introduction; in Section 3.6, I will prove this claim). So, this condition is philosophically quite interesting.

$C- \mathcal{W}C$ is another philosophically interesting condition. According to $C- \mathcal{W}C$, for every (individual) d, for every (possible world) x there is

[17] '$M\mathcal{W}$' in '$C- M\mathcal{W}$' stands for 'Must Want,' and '$\mathcal{W}C$' in '$C- \mathcal{W}C$' for 'Want Can.' $C- ab4$ (as in 'alethic boulesic 4') is called '$C- ab4$' because it is similar to the well-known alethic (modal) condition $C- 4$ and the alethic deontic condition $C- ad4$, and similarly for $C- ad5$. '$\mathcal{A}MP$' in '$C- \mathcal{A}MP$' is an abbreviation of 'Acceptance Must Permutation,' and '$\mathcal{W}MP$' in '$C- \mathcal{W}MP$' is an abbreviation of 'Want Must Permutation.'

a (possible world) y such that d accepts y in x and y is alethically accessible from x. In other words, in every possible world, d accepts at least one possible world that is alethically accessible. This condition is similar to condition $C-bD$ (Table 3). $C-\mathcal{W}C$ entails $C-bD$, but $C-bD$ (in itself) does not entail $C-\mathcal{W}C$. In every class of models that satisfies this condition, the following schema is valid: $\Pi x(Rx \rightarrow (\mathcal{W}_x A \rightarrow \Diamond A))$ ('For every x: if x is perfectly rational, then x wants it to be the case that A only if A is (historically) possible'). Hence, according to this condition, a perfectly rational individual does not want anything impossible. This is an intuitively plausible principle. If c wants something that is impossible, c's want will inevitably be frustrated; c can never see to it that anything impossible is the case. In the introduction, we called this principle the 'Want-Can principle' ($\mathcal{W}C$).

Space does not allow me to discuss every philosophically interesting argument for and against the Want-Can principle, but I would like to address one possible counterexample (this problem was raised by an anonymous reviewer). If $\mathcal{W}C$ is valid, then we must accept the following instance of this principle: If a perfectly rational individual wants a unicorn to exist then it is (historically) possible that a unicorn exists. But this instance is counterintuitive. Hence, $\mathcal{W}C$ cannot be valid. I agree that this instance seems somewhat strange, even perhaps counterintuitive, at a first glance. But I do not think that this fact refutes the principle. Let me explain why. The expression 'If, then' in this instance should be interpreted as material implication. Sometimes we read more into this expression in English, for example, some kind of causal relation. But we should avoid this in our case. Suppose that a perfectly rational individual c wants it to be the case that A. $\mathcal{W}C$ does not entail that we have to assume that c's attitude *causes* it to be the case that it is historically possible that A. If we are idealists about possibilities, we might want to make such an assumption. But we do not have to be idealists to accept the systems in this paper. We *can* read $\mathcal{W}C$ in the following way: 'For every x: if x is perfectly rational, then *if* x wants it to be the case that A *then* A is (historically) possible.' But, as I suggested above, it might be more plausible to read it in the other direction, that is, in the following way: 'For every x: if x is perfectly rational, then x wants it to be the case that A *only if* A is (historically) possible.' We can think of our example in this way. Then we should read our instance of $\mathcal{W}C$ in the following way: ‘If someone is perfectly rational, then she wants a unicorn to exist only if it is (historically) possible that a unicorn exists.’ Suppose that it is not (historically) possible that a unicorn exists. (In our possible world it seems to be the case that it is logically but *not* historically possible that there exists a unicorn.) Then a perfectly rational individual will adjust his or her attitudes to this fact. Hence, he or she will not want a unicorn to exist. Furthermore, recall that we use ‘want’ in an all-things-considered sense in this paper. So, the fact that a perfectly rational individual does not want anything that is (historically) impossible according to $\mathcal{W}C$ does not necessarily entail that he or she cannot daydream about unicorns, think about what it would be like if a unicorn existed, believe that it would be cool if a unicorn existed, etc. (see footnote 3). However, he or she will not want a unicorn to exist in an all-things-considered sense. So, I do not think that this example refutes $\mathcal{W}C$.

3.3.6. Conditions concerning the relation between $\mathfrak{S}$ and $\mathfrak{A}$

Condition	Formalisation of condition
$C-O\mathcal{W}$	$\forall x \forall y(\exists d\mathfrak{A}dxy \rightarrow \mathfrak{S}xy)$
$C-\mathcal{W}O$	$\forall x \forall y(\mathfrak{S}xy \rightarrow \exists d\mathfrak{A}dxy)$
$C-\mathfrak{A}\Sigma$	$\forall x \forall y(\mathfrak{S}xy \leftrightarrow \exists d\mathfrak{A}dxy)$
$C-O\mathcal{W}'$	$\forall x \forall y(\forall d\mathfrak{A}dxy \rightarrow \mathfrak{S}xy)$
$C-\mathcal{W}O'$	$\forall x \forall y(\mathfrak{S}xy \rightarrow \forall d\mathfrak{A}dxy)$
$C-\mathfrak{A}\Pi$	$\forall x \forall y(\mathfrak{S}xy \leftrightarrow \forall d\mathfrak{A}dxy)$

Table 6[18]

The conditions in Table 6 are concerned with some possible relationships between the deontic accessibility relation $\mathfrak{S}$ and the boulesic accessibility relation $\mathfrak{A}$. According to $C-\mathfrak{A}\Sigma$, y is deontically accessible from x iff y is acceptable to at least one individual in x; and according to $C-\mathfrak{A}\Pi$, y is

[18] ‘$O\mathcal{W}$’ is an abbreviation of ‘Ought Want,’ and ‘$\mathcal{W}O$’ of ‘Want Ought.’

deontically accessible from x iff y is acceptable to all individuals in x. $C-\mathfrak{A}\Pi$ is an immediate consequence of ($Def\, \mathfrak{S}$), which is a definition that we will introduce in Section 3.4. $C-\mathfrak{A}\Sigma$ follows from $C-O\mathcal{W}$ and $C-\mathcal{W}O$, and $C-\mathfrak{A}\Pi$ follows from $C-O\mathcal{W}'$ and $C-\mathcal{W}O'$. If the condition $C-HW$ (Table 7) holds, then $C-\mathfrak{A}\Sigma$ and $C-\mathfrak{A}\Pi$ are equivalent, for then a possible world y is boulesically accessible from a possible world x to some individual iff y is boulesically accessible from x to every individual. $C-O\mathcal{W}$ corresponds to the tableau rule $T-O\mathcal{W}$ and $C-\mathcal{W}O$ to the tableau rule $T-\mathcal{W}O$ (Table 19). In Sections 3.4 and 4.2.12, we will consider some consequences of $C-\mathfrak{A}\Pi$ ($Def\, \mathfrak{S}$). In every model that satisfies $C-O\mathcal{W}$ the following sentence is valid: $OA \to \Pi x(Rx \to \mathcal{W}_x A)$, which says that if it ought to be the case that A then everyone who is perfectly rational wants it to be the case that A. If a model satisfies $C-\mathcal{W}O$ (and $C-HW$ and $C-\Sigma R$ in Table 7), then $\Pi x(Rx \to \mathcal{W}_x A) \to OA$ is valid in this model. $\Pi x(Rx \to \mathcal{W}_x A) \to OA$ says that it ought to be the case that A if everyone who is perfectly rational wants it to be the case that A. The intuitive idea behind the conditions in Table 6 is that there might be some interesting connections between what perfectly rational individuals want and accept and various normative 'facts' about what ought to be the case, about what is permitted and about what is not permitted. In our systems, we can explore those possible connections in a systematic and precise way. (For more on this, see Sections 3.4 and 4.2.12.)

3.3.7. One more condition on the relation $\mathfrak{A}$ and the being of a perfectly rational individual

Condition	Formalisation of condition
$C-HW$	$\forall c \forall d \forall x \forall y(\mathfrak{A}cxy \to \mathfrak{A}dxy)$
$C-\Sigma R$	In every possible world, ω, there is at least one individual, d, such that d is in R's extension in ω.

Table 7[19]

[19] 'HW' in '$C-HW$' is an abbreviation of '[the] Harmony of the Wills' and 'R' in '$T-\Sigma R$' is an abbreviation of 'perfectly rational' or 'perfectly reasonable.'

$C\!-HW$ says that if the possible world y is acceptable to the individual c in the possible world x, then y is acceptable to any individual d in x. $C\!-HW$ corresponds to the tableau rule $T\!-HW$ (Section 4.2.8, Table 15).

$C\!-\Sigma R$ says the following: in every possible world, there is at least one individual that is perfectly rational. This does not entail that there is one individual that is perfectly rational in every possible world. However, if there is one individual that is perfectly rational in every possible world, then obviously $C\!-\Sigma R$ holds. $C\!-\Sigma R$ corresponds to the tableau rule $T\!-\Sigma R$ (Section 4.2.8, Table 15). When I say that 'there is at least one individual,' I do not mean that this necessarily entails that this individual exists in this world. The expression is interpreted as a kind of 'possibilist quantifier' that is supposed to range over every possible object. Obviously, though, if there exists an individual that is perfectly rational (in some possible world), then there is such an individual (in this world).

3.3.8. Conditions on the valuation function *v* in a model

Condition	Formalisation of condition
$C\!-FTR$	If $\mathfrak{R}\omega_1\omega_2$ and Rc is true in ω_1, then Rc is true in ω_2 (for any c).
$C\!-UR$	If Rc is true in ω_1, then Rc is true in ω_2 (for any c).

Table 8

The semantic conditions $C\!-FTR$ and $C\!-UR$ correspond to the tableau rules $T\!-FTR$ and $T\!-UR$, respectively. (See Section 4.2.10, Table 17, for more on this.)

3.4. Relations between semantic conditions

There are many interesting relationships between the conditions introduced in Section 3.3. It is not possible to go through them all, but I will mention some of the most interesting. Due to considerations of space, proofs are omitted. Let us begin by saying a few words about the alethic accessibility relation.

Remark 2 *The following facts are well-known. (a) If $\mathfrak{R}$ is reflexive ($C\!-aT$), then $\mathfrak{R}$ is serial ($C\!-aD$). (b) $\mathfrak{R}$ is an equivalence relation iff (i) $\mathfrak{R}$ is reflexive ($C\!-aT$), symmetric ($C\!-aB$) and transitive ($C\!-a4$), iff (ii) $\mathfrak{R}$ is reflexive ($C\!-aT$) and Euclidean ($C\!-a5$), iff (iii) $\mathfrak{R}$ is serial ($C\!-aD$), symmetric ($C\!-aB$) and transitive ($C\!-a4$), iff (iv) $\mathfrak{R}$ is serial ($C\!-aD$), symmetric ($C\!-aB$) and Euclidean ($C\!-a5$).*

Since $\Box$ and $\Diamond$ are interpreted as historical necessity and possibility, it is reasonable to treat $\mathfrak{R}$ as an equivalence relation. If we assume this, $\Box$ and $\Diamond$ will behave as so-called $S5$-operators.

The following theorem says something about the relationships between the conditions on the deontic accessibility relation and between the conditions on the boulesic accessibility relation.

Remark 3 *The following facts are well-known. (i) If $\mathfrak{S}$ is Euclidean ($C\!-d5$), then $\mathfrak{S}$ is almost (secondarily) reflexive ($C\!-dT'$) and almost (secondarily) symmetric ($C\!-dB'$). Likewise, it is easy to prove the following facts. (ii) If $\mathfrak{A}$ is Euclidean ($C\!-b5$), then $\mathfrak{A}$ is almost (secondarily) reflexive ($C\!-bT'$) and almost (secondarily) symmetric ($C\!-bB'$).*

Now we will introduce some definitions. First, we define the ternary boulesic accessibility relation $\mathfrak{A}$ and the binary deontic accessibility relation $\mathfrak{S}$, and then we investigate the consequences of these definitions. We also introduce a new semantic condition called *the Accessibility Condition*, or $C\!-bdD$ (Definition 6), which has some interesting implications.[20]

Definition 4 *Def* $\mathfrak{A}$. $\forall c\forall x\forall y(\mathfrak{A}cxy =_{df} (\mathfrak{R}xy \wedge Tcy))$.[21]

In this definition (Definition 4), T is a binary predicate that says that (the individual) c accepts (the possible world) y or that y is acceptable to c. *Def* $\mathfrak{A}$ can be read as: 'For every (individual) c and for all (possible worlds) x and y: y is acceptable to c in x iff y is alethically accessible from

[20] Note that the definitions in this paper are 'theoretical' rather than 'lexical' or 'descriptive.' Furthermore, the 'definiens' does not have to have the same meaning as the 'definiendum' in every respect.

[21] Note that this definition does not entail any of the following propositions: $\forall c\exists yTcy$, $\exists c\exists yTcy$, $\forall c\exists y(Rcy \wedge Tcy)$, $\exists c\exists y(Rcy \wedge Tcy)$.

x and c accepts y.' This is a definition of the ternary boulesic accessibility relation $\mathfrak{A}$ in terms of the alethic accessibility relation $\mathfrak{R}$ and the binary accessibility relation T.[22]

Definition 5 *Def* $\mathfrak{S}$. $\forall x \forall y(\mathfrak{S}xy =_{df} \forall c \mathfrak{A}cxy)$.

Def $\mathfrak{S}$ (Definition 5) was mentioned already in Section 3.3.6. It should be obvious that $C-\mathfrak{A}\Pi$ is an immediate consequence of *Def* $\mathfrak{S}$. *Def* $\mathfrak{S}$ is a definition of the deontic accessibility relation $\mathfrak{S}$ in terms of the boulesic accessibility relation $\mathfrak{A}$. Informally, Definition 5 says the following: 'For all (possible worlds) x and y: y is deontically accessible from x iff every (individual) c accepts y in x' (or iff y is acceptable to every individual c in x). The intuition behind this definition is that the aim of morality is to create a possible world that everyone (who is perfectly rational) accepts (or can accept). This is an idea that might be attractive to at least some ideal observer theorists, Kantians, contractualists, moral idealists, constructivists, response dependent theorists, and divine will theorists. The definition has several interesting formal consequences (see Theorem 7 and Theorem 8 below), but it does not tell us anything about which worlds various

[22] *Def* $\mathfrak{A}$ is compatible with many different theories about what it means for an individual to accept a possible world (for a world to be acceptable to an individual). T is not necessarily a primitive, undefined relation. Here are some possible definitions: y is acceptable to c iff the utility of y for c is positive, or above a certain threshold or as high as possible, or iff c does not prefer any other possible world to y, or . . . . The important thing for our purposes in this paper is that all definitions of this kind share the same form. The definition is also consistent with the proposition that different individuals accept different worlds and that different individuals might have different reasons for accepting a possible world. Perhaps c accepts y because the utility of y for c is above a certain threshold, and perhaps d accepts z because z does not contain any serious violations of human rights. It is an interesting question whether or not T is definable, but for our purposes in this paper, we do not have to answer this question. However, note that not all definitions are compatible with condition $C-bdD$ (see definition 6) or with the proposition that there are possible worlds that are acceptable to some individual in some possible world. Suppose, for example, that y is acceptable to c iff the utility of y for c is positive and that there are no alethically accessible worlds from the world w in which the utility of y for c is positive. Then there is no world in w that is acceptable to c.

individuals accept or why they accept them (see footnote 22). So, the view is compatible with several different value theories and substantive normative theories.

Definition 5 does not entail that there is any world that everyone (who is perfectly rational) accepts. Accordingly, this definition does not guarantee that ought implies historical possibility. To guarantee this principle, we need to introduce another condition, namely the following:

Definition 6 *(C− bdD) The Accessibility Condition.* $\forall x \exists y (\mathfrak{R}xy \wedge \forall c Tcy)$.

The Accessibility Condition says the following: 'For every (possible world) x, there is a (possible world) y such that y is alethically accessible from x and for every (individual) c: c accepts y'; in other words, in every possible world there is at least one alethically accessible world that everyone accepts (or that is acceptable to everyone). Intuitively, this condition entails that no matter how good or bad things are in a given situation (possible world) everyone (who is perfectly rational) will agree that at least one possible outcome (world) is acceptable in this situation (world).[23] This definition (together with some other conditions) guarantees that ought implies historical possibility, that is, if we assume this condition, then everything that ought to be is historically possible and nothing historically impossible is obligatory (see Theorem 8 below).

Now we will investigate some consequences of these definitions.

Theorem 7 *Suppose that $\mathfrak{S}$ can be defined in terms of $\mathfrak{A}$ according to Definition 5 (Def $\mathfrak{S}$) and that the Harmony of the Wills holds (C− HW). Furthermore, suppose that $\mathfrak{A}$ is serial (C− bD), transitive (C− b4) and Euclidean (C− b5). Then $\mathfrak{A}$ is almost reflexive (C− bT′) and almost symmetric*

[23] I do not suggest that The Accessibility Condition means that everyone '*consciously*' accepts at least one possible world. Ordinary people almost certainly do not have any conscious attitudes that involve whole possible worlds and they disagree about many things. But $\mathfrak{A}$ is only relevant for perfectly rational individuals; the truth values of sentences of the forms $\mathcal{W}_cB$ and $\mathcal{A}_cB$ when c is not perfectly rational do not depend on $\mathfrak{A}$. We are primarily interested in the structure of a perfectly rational will, not about the actual attitudes of ordinary people; T is an 'ideal' relation. So, this is not a problem for The Accessibility Condition.

($C\!-bB'$), and $\mathfrak{S}$ is serial ($C\!-dD$), transitive ($C\!-d4$), Euclidean ($C\!-d5$) and (hence) almost reflexive ($C\!-dT'$) and almost symmetric ($C\!-dB'$).

Theorem 8 *Suppose $\mathfrak{R}$ is an equivalence relation. (i) Then Def $\mathfrak{A}$ and Def $\mathfrak{S}$ entail the following conditions: $d4$, $d5$ and hence dT' and dB', $b4$, $b5$ and hence bT' and bB', MO, OC′, MO′, ad4, ad5, OMP, PMP, $M\mathcal{W}$, $\mathcal{W}C'$, $M\mathcal{W}'$, bd4, bd5, $\mathcal{W}MP$ and $\mathcal{A}MP$. (ii) Then Def $\mathfrak{A}$, Def $\mathfrak{S}$ and HW entail all conditions in (i), and all conditions in Table 6. (iii) Then Def $\mathfrak{A}$, Def $\mathfrak{S}$ and $C\!-bdD$ entail all the conditions in tables 1–5. (iv) Then Def $\mathfrak{A}$, Def $\mathfrak{S}$, HW and $C\!-bdD$ entail all the conditions in tables 1–6.*

Accordingly, if $C\!-HW$ and $C\!-bdD$ are plausible and the definitions in this section are reasonable (they do have significant intuitive appeal), then we have a good reason to accept all conditions in tables 1–6. The conditions in Table 7 and Table 8 might seem controversial. Nevertheless, I think one could make a good case for accepting them (at least every condition except $C\!-UR$[24]). However, space does not permit me to do this in the present paper. Whether or not we should accept all conditions in this section, clearly all of them are interesting enough to be worth discussing. (See sections 4.2.8 and 4.2.10 for more on some tableau rules that correspond to the semantic conditions in tables 7 and 8.)

3.5. Model classes and the logic of a class of models

The conditions mentioned in Section 3.3 can be used to obtain a categorisation of the set of all models into various kinds. We shall say that $\mathcal{M}(C_1, \ldots, C_n)$ is the class of (all) models that satisfy the conditions

[24] $C\!-UR$ is a theoretically important condition. Yet, there might be good reasons to reject it. Even though we, human beings, are not perfectly rational (see footnote 6), it seems interesting to consider what would be the case if we were. If all perfectly rational individuals necessarily are perfectly rational, we cannot do this, for then there are no individuals that are contingently perfectly rational—i.e. perfectly rational in some possible worlds and not perfectly rational in some other possible worlds. A being that is in fact not perfectly rational cannot then be perfectly rational in some other possible world.

$C_1, \ldots, C_n$. For example, $\mathcal{M}(C-bD, C-b4, C-b5)$ is the class of (all) models that satisfy the conditions $C-bD$, $C-b4$ and $C-b5$.

By imposing different conditions on our models we can obtain different logical systems. The set of all sentences in the language $L1$ ($L2$) that are valid in a class of models $\mathcal{M}$ is called the logical system of $\mathcal{M}$, or the system of $\mathcal{M}$, or the logic of $\mathcal{M}$, in symbols $\mathcal{S}(\mathcal{M})$. For example, $\mathcal{S}(\mathcal{M}(C-bD, C-b4, C-b5))$ (the system of $\mathcal{M}(C-bD, C-b4, C-b5)$) is the class of sentences in $L1$ ($L2$) that are valid in the class of (all) models that satisfy the conditions $C-bD$, $C-b4$ and $C-b5$.

By using this classification of model classes we can define a large set of systems. In the next section, I will develop semantic tableau systems that exactly correspond to these logics. I will consider four systems that seem especially philosophically interesting. The first is a pure boulesic system; the other three are boulesic-deontic systems (Section 4.3).

Definition 9 *(i) Let the class of all* strict models *be the class of models where $\mathfrak{R}$ is an equivalence relation and where Def $\mathfrak{A}$ holds. (ii) Let the class of all* strong models *be the class of all (supplemented) models where $\mathfrak{R}$ is an equivalence relation and where Def $\mathfrak{A}$, Def $\mathfrak{S}$, $C-HW$ and $C-\Sigma R$ hold. (iii) Let the class of all* strong+ models *be the class of all (supplemented) models where $\mathfrak{R}$ is an equivalence relation and where Def $\mathfrak{A}$, Def $\mathfrak{S}$, $C-bdD$, $C-HW$ and $C-\Sigma R$ hold. (iv) Let the class of all* almost complete models *be the class of all (supplemented) models where $\mathfrak{R}$ is an equivalence relation and where Def $\mathfrak{A}$, Def $\mathfrak{S}$, $C-bdD$, $C-HW$, $C-\Sigma R$ and $C-FTR$ hold.*

The first class in Definition 9 corresponds to strict boulesic logic, the second to strong boulesic-deontic logic, the third to strong+ boulesic-deontic logic, and the fourth to almost complete boulesic-deontic logic (Section 4.3, Definition 10). Hence, the system of the class of all strict models is the same as the set of all sentences provable in strict boulesic logic (see Section 4.3, Definition 10), etc. This follows from the soundness and completeness results in Section 6 and the results in Section 3.4.

3.6. An example of a valid formula

In the introduction, I mentioned the so-called hypothetical imperative. One of the most interesting readings of this principle was interpretation (6): $U\Pi x(Rx \to ((\mathcal{W}_x A \wedge \Box(A \to B)) \to \mathcal{W}_x B))$. Now I will show that this formula is valid in the class of all models that satisfy C–$M\mathcal{W}$ (Table 5).[25]

To establish this, assume that this sentence is not true in some possible world ω in some model $\mathcal{M}$ that satisfies $C-M\mathcal{W}$. Then there is some possible world ω' in $\mathcal{M}$ in which $\Pi x(Rx \to ((\mathcal{W}_x A \wedge \Box(A \to B)) \to \mathcal{W}_x B))$ is false. Hence, Rc, $\mathcal{W}_c A$ and $\Box(A \to B))$ are true in ω' in $\mathcal{M}$, while $\mathcal{W}_c B$ is false in ω' in $\mathcal{M}$ (‘c’ represents an arbitrary ‘new’ individual). Since c is perfectly rational in ω' in $\mathcal{M}$ and $\mathcal{W}_c B$ is false in ω' in $\mathcal{M}$, there is a possible world ω'' in $\mathcal{M}$ that is boulesically accessible to c from ω' in which B is false. c is perfectly rational in ω', ω'' is boulesically accessible to c from ω' and $\mathcal{W}_c A$ is true in ω' in $\mathcal{M}$. Hence, A is true in ω'' in $\mathcal{M}$. Since $\mathcal{M}$ satisfies $C-M\mathcal{W}$ and ω'' is boulesically accessible to c from ω' in $\mathcal{M}$, ω'' is alethically accessible from ω' in $\mathcal{M}$. Consequently, $A \to B$ is true in ω'' in $\mathcal{M}$, for ω'' is alethically accessible from ω' and $\Box(A \to B)$ is true in ω' in $\mathcal{M}$. Therefore, B is true in ω'' in $\mathcal{M}$ (by propositional logic). But this is absurd. Accordingly, our assumption cannot be true. In conclusion, $U\Pi x(Rx \to ((\mathcal{W}_x A \wedge \Box(A \to B)) \to \mathcal{W}_x B))$ is valid in $\mathcal{M}$. Since, ω and $\mathcal{M}$ were arbitrary, it follows that $U\Pi x(Rx \to ((\mathcal{W}_x A \wedge \Box(A \to B)) \to \mathcal{W}_x B))$ is valid in every model that satisfies $C-M\mathcal{W}$. Q.E.D.

4. Proof theory

4.1. Semantic tableaux

In Section 4, I will develop a set of tableau systems. The propositional part of these systems is similar to systems introduced by Raymond

[25] In a strict sense, $U\Pi x(Rx \to ((\mathcal{W}_x A \wedge \Box(A \to B)) \to \mathcal{W}_x B))$ is not a sentence but a schema. The argument in this section shows that every instance of this schema is valid in the class of all models that satisfy $C-M\mathcal{W}$.

Smullyan (1968) and Richard Jeffrey (1967), and the modal part is similar to systems discussed by Graham Priest (2008). For more information about the tableau method and various kinds of tableau systems, see, for example, (D'Agostino et al. 1999; and Fitting and Mendelsohn 1998).

The concepts of semantic tableau, branch, open and closed branch, etc. are essentially defined as in (Priest 2008).

4.2. Tableau rules

In this section, I will introduce a set of tableau rules that can be used to construct a large set of tableau systems (Section 4.3). They should be more or less self-explanatory. However, I will comment on some of the new rules.

4.2.1. Propositional rules

I will use the same propositional rules as in (Priest 2008). Let us call them $(\neg\neg)$, $(\wedge)$, $(\neg\wedge)$, $(\vee)$, $(\neg\vee)$, $(\rightarrow)$, $(\neg\rightarrow)$, $(\leftrightarrow)$ and $(\neg\leftrightarrow)$.

4.2.2. Basic alethic rules (ba-rules)

U	M	$\Box$	$\Diamond$
UA,i $\downarrow$ A,j for any j	MA,i $\downarrow$ A,j where j is new	$\Box A,i$ irj $\downarrow$ A,j	$\Diamond A,i$ $\downarrow$ irj A,j where j is new
$\neg U$	$\neg M$	$\neg\Box$	$\neg\Diamond$
$\neg UA,i$ $\downarrow$ $M\neg A,i$	$\neg MA,i$ $\downarrow$ $U\neg A,i$	$\neg\Box A,i$ $\downarrow$ $\Diamond\neg A,i$	$\neg\Diamond A,i$ $\downarrow$ $\Box\neg A,i$

Table 9

4.2.3. Basic boulesic and deontic rules (bb-rules and d-rules)

$\mathcal{W}$	$\mathcal{A}$	O	P
Rc,i $\mathcal{W}_cB,i$ $iAcj$ $\downarrow$ B,j	Rc,i $\mathcal{A}_cB,i$ $\downarrow$ $iAcj$ B,j where j is new	OB,i isj $\downarrow$ B,j	PB,i $\downarrow$ isj B,j where j is new
$\neg\mathcal{W}$	$\neg\mathcal{A}$	$\neg O$	$\neg P$
Rc,i $\neg\mathcal{W}_cB,i$ $\downarrow$ $\mathcal{A}_c\neg B,i$	Rc,i $\neg\mathcal{A}_cB,i$ $\downarrow$ $\mathcal{W}_c\neg B,i$	$\neg OB,i$ $\downarrow$ $P\neg B,i$	$\neg PB,i$ $\downarrow$ $O\neg B,i$

Table 10

Intuitively, 'Rc,i' in the boulesic rules says that the individual denoted by 'c' is perfectly rational in the possible world denoted by 'i,' and '$iAcj$' in the rules $\mathcal{W}$ and $\mathcal{A}$ says that the possible world denoted by 'j' is acceptable to the individual denoted by 'c' in the possible world denoted by 'i.' The basic boulesic rules hold for every constant c (i.e. c can be replaced by any constant in these rules).

4.2.4. Possibilist quantifiers

Π	Σ	$\neg\Pi$	$\neg\Sigma$
$\Pi xA,i$ $\downarrow$ $A[a/x],i$ for every constant a on the branch, a new if there are no constants on the branch	$\Sigma xA,i$ $\downarrow$ $A[c/x],i$ where c is new to the branch	$\neg\Pi xA,i$ $\downarrow$ $\Sigma x\neg A,i$	$\neg\Sigma xA,i$ $\downarrow$ $\Pi x\neg A,i$

Table 11

Note that a and c in the quantifier rules are rigid constants—we never instantiate with variables; a is any constant on the branch and c is a constant new to the branch.

4.2.5. Alethic accessibility rules (a-rules)

$T - aD$	$T - aT$	$T - aB$	$T - a4$	$T - a5$
i $\downarrow$ irj where j is new	i $\downarrow$ iri	irj $\downarrow$ jri	irj jrk $\downarrow$ irk	irj irk $\downarrow$ jrk

Table 12

4.2.6. Deontic accessibility rules (d-rules)

$T - dD$	$T - d4$	$T - d5$	$T - dT'$	$T - dB'$
i $\downarrow$ isj where j is new	isj jsk $\downarrow$ isk	isj isk $\downarrow$ jsk	isj $\downarrow$ jsj	isj jsk $\downarrow$ ksj

Table 13

4.2.7. Boulesic accessibility rules (b-rules)

$T - bD$	$T - b4$	$T - b5$	$T - bT'$	$T - bB'$
i $\downarrow$ $iAcj$ where j is new	$iAcj$ $jAck$ $\downarrow$ $iAck$	$iAcj$ $iAck$ $\downarrow$ $jAck$	$iAcj$ $\downarrow$ $jAcj$	$iAcj$ $jAck$ $\downarrow$ $kAcj$

Table 14

The boulesic accessibility rules hold for every constant c (i.e. c can be replaced by any constant in these rules). The b-rules in Table 14 correspond to the semantic conditions in Table 3.

4.2.8. Non-basic boulesic rules (nbb-rules)

$T-HW$	$T-\Sigma R$
$iAcj$ $\downarrow$ $iAdj$ for every c and d	i $\downarrow$ $\Sigma xRx,i$

Table 15

The rules in Table 15 correspond to the semantic conditions in Table 7. In every system that includes $T-HW$, we can prove the first six sentences in Table 25, and in every system that includes $T-HW$ and $T-bD$, we can prove all sentences in Table 25 (see Section 5). Therefore, in systems that include $T-HW$, we can prove that all perfectly rational individuals want and accept the same things. According to this rule, the idea of perfect rationality, or wisdom, includes a kind of *inter*personal consistency, not only a kind of *intra*personal consistency. The wills of perfectly rational individuals are consistent, they harmonise. If individual c wants it to be the case that B and individual d wants it to be the case that $\neg B$, then both cannot get what they want; either c's or d's desires will be frustrated: it is not possible to see to it that B and to see to it that $\neg B$. In systems that include $T-HW$ and $T-bD$, situations of this kind are ruled out. Hence, these conditions seem to be intuitively plausible (however, see footnote 14).

If we include $T-\Sigma R$ in our systems, we can prove that ΣxRx is necessarily true. Recall that ΣxRx says that there is something or someone, a possible individual that is perfectly rational. This does not entail that this individual exists.

Space does not permit me to discuss all philosophical arguments for and against these rules. However, it should be noted that $T-HW$ does not entail that all individuals that are *not* perfectly rational want and accept the same things, and it does not entail that everyone *should* have the same attitudes. Furthermore, it does not entail that everyone should act in the same way or be a certain kind of person, nor does it entail that if something is permitted for some person it is permitted for every person. Suppose that c and

d are perfectly rational. Even if this is the case, it is possible that both c and d want individual e to perform a certain action and that both want individual f not to perform this action. Situations of this kind are not inconsistent according to any system in this paper. If they were inconsistent, $T-HW$ would probably not be a philosophically reasonable rule.

4.2.9. The CUT-rule (CUT)

CUT
i
$\swarrow \ \searrow$
$A{,}i \quad \neg A{,}i$
for every A and i

Table 16[26]

4.2.10. Transfer-rules, etc.

$T-FTR$	$T-UR$
$Rc{,}i$	$Rc{,}i$
irj	$\downarrow$
$\downarrow$	$Rc{,}j$
$Rc{,}j$	for any j

Table 17[27]

The tableau rules in Table 17 correspond to the semantic conditions in Table 8.

[26] We could use a more restricted CUT rule, $CUTR$, where 'A' in CUT is replaced by 'Rc' where c is a constant (that occurs as an index to some boulesic operator) on the branch. In fact, in the completeness proofs we do not need CUT if our systems include $CUTR$. However, CUT is often more useful in proving theorems and deriving non-primitive rules. For more on the CUT rule, see, for example, (Rönnedal 2009).

[27] 'FT' in '$T-FTR$' is an abbreviation of 'Forward Transfer,' and 'R' in '$T-FTR$' and '$T-UR$' of 'Rationality.'

In every system that includes $T-UR$ or $T-FTR$ and $T-M\mathcal{W}$ (Table 18), we can prove that the following sentence is a theorem: $\Pi x(Rx \rightarrow \mathcal{W}_x Rx)$, which says that everyone who is perfectly rational wants to be perfectly rational.

In every system that includes $T-UR$ or $T-FTR$ and $T-M\mathcal{W}$, and $T-bD$ (Table 14), we can prove that the following sentence is a theorem: $\Pi x(Rx \rightarrow \mathcal{A}_x Rx)$, which says that everyone who is perfectly rational accepts that she is perfectly rational.

In every system that includes $T-UR$, we can prove the following sentence: $\Pi x(Rx \rightarrow URx)$, which says that every perfectly rational individual is necessarily perfectly rational.

We do not assume that the transfer rules (the rules in Table 17) are included in every system. Whether or not they should be added seems to be something of an open question.[28]

4.2.11. Alethic-boulesic accessibility rules (ab-rules)

$T-M\mathcal{W}$	$T-M\mathcal{W}'$	$T-\mathcal{W}C$	$T-\mathcal{W}C'$
$iAcj$ $\downarrow$ irj	$iAcj$ $jAck$ $\downarrow$ jrk	i $\downarrow$ $iAcj$ irj where j is new	$iAcj$ $\downarrow$ $jAck$ jrk where k is new
$T-ab4$	$T-ab5$	$T-\mathcal{A}MP$	$T-\mathcal{W}MP$
irj $jAck$ $\downarrow$ $iAck$	irj $iAck$ $\downarrow$ $jAck$	$iAcj$ irk $\downarrow$ jrl $kAcl$ where l is new	irj $jAck$ $\downarrow$ $iAcl$ lrk where l is new

Table 18

[28] See footnote 24 for some critique of $C-UR$, which is the semantic condition that corresponds to $T-UR$.

The ab-rules in Table 18 correspond to the semantic conditions introduced in Table 5.

4.2.12. Boulesic-deontic accessibility rules (bd-rules)

$T-O\mathcal{W}$	$T-\mathcal{W}O$
$iAbj$	isj
$\downarrow$	$\downarrow$
isj	$iAcj$
for any b	where c is new

Table 19

The rule $T-O\mathcal{W}$ in Table 19 corresponds to the semantic condition $C-O\mathcal{W}$ mentioned in Table 6, and the rule $T-\mathcal{W}O$ in Table 19 corresponds to the semantic condition $C-\mathcal{W}O$ in Table 6. In every system that includes $T-O\mathcal{W}$, we can prove that $OA \rightarrow \Pi x(Rx \rightarrow \mathcal{W}_xA)$ is a theorem—i.e. if it ought to be the case that A, then everyone who is perfectly rational wants it to be the case that A. This is one version of a philosophically very interesting thesis often called 'existence internalism.' It follows from this theorem that if the individual c ought to do the action H, then if c is perfectly rational c wants to do H. However, if c is not perfectly rational, it is not necessary that she wants to do H. So, this kind of internalism is compatible with the existence of amoralists and with the phenomenon of weakness of will. Internalism can help explain the fact that we find utterances of the following kind puzzling: 'I know that I ought to do it, but I have no inclination whatsoever to do it' and 'You ought to do it, but by all means don't do it.' At the same time, the kind of internalism mentioned here avoids some of the common objections against this thesis.[29]

[29] For more information on internalism and various versions of internalism and arguments for and against this thesis, see, for example, (Björklund et.al. 2012; Björnsson et.al. 2015; and van Roojen 2013). It might be interesting to note that existence internalism entails the following version of 'knowledge internalism': $\mathcal{K}_cOA \rightarrow (Rc \rightarrow \mathcal{W}_cA)$, where '$\mathcal{K}_cA$' stands for '$c$ knows that A' (given that knowledge implies truth). '$\mathcal{K}_cOA \rightarrow (Rc \rightarrow \mathcal{W}_cA)$' says that if c knows that it ought to be the case that A, then if c is perfectly rational then c wants it to be the case that A.

In every system that includes $T-\mathcal{W}O$, $T-H\mathcal{W}$ and $T-\Sigma R$ we can prove that $\Pi x(Rx \to \mathcal{W}_x A) \to OA$ is a theorem—i.e. if everyone who is perfectly rational wants it to be the case that A, then it ought to be the case that A. This is the converse of $OA \to \Pi x(Rx \to \mathcal{W}_x A)$. Together these theorems entail $OA \leftrightarrow \Pi x(Rx \to \mathcal{W}_x A)$, which says that it ought to be the case that A iff everyone who is perfectly rational wants it to be the case that A. Similar equivalences hold for P (it is permitted that) and F (it is not permitted that) (see Table 30). These theorems can be seen as a part of a kind of ideal observer theory for normative propositions.[30]

Some might worry that the equivalences in Table 30 are too strong. If we accept those equivalences, do we not have to accept that, for example, 'Tom ought to go home' has the same meaning as 'Everyone who is perfectly rational wants Tom to go home' and isn't this unreasonable?[31] Personally, I do not think that we have to accept this. Let me explain why. Since $OA \leftrightarrow \Pi x(Rx \leftrightarrow \mathcal{W}_x A)$ holds in some systems, OA is in principle 'definable' in terms of $\Pi x(Rx \to \mathcal{W}_x A)$ in those systems (see Section 2.3). However, this fact does not entail that 'It ought to be the case that A' has the same meaning as 'Everyone who is perfectly rational wants it to be the case that A.' To say that OA is in principle 'definable' in terms of $\Pi x(Rx \to \mathcal{W}_x A)$ means that 'OA' can be replaced by '$\Pi x(Rx \to \mathcal{W}_x A)$' (and vice versa) in every 'extensional context,' but not necessarily in every 'intensional context,' for example, if 'OA' ('$\Pi x(Rx \to \mathcal{W}_x A)$') occurs within the scope of a boulesic operator. So, those systems do not entail that, for example, 'Tom ought to go home' says exactly the same thing as 'Everyone who is perfectly rational wants Tom to go home.' In this sense, our equivalences are similar

Furthermore, assume that every perfectly rational individual is infallible in the sense that everything she believes is true. Then existence internalism entails the following version of 'belief internalism': $\mathcal{B}_c OA \to (Rc \to \mathcal{W}_c A)$, where '$\mathcal{B}_c A$' stands for '$c$ believes that A.' '$\mathcal{B}_c OA \to (Rc \to \mathcal{W}_c A)$' says that if c believes that it ought to be the case that A, then if c is perfectly rational then c wants it to be the case that A.

[30] For more on ideal observer theories, see, for example, (Firth 1952; and Kawall 2013).

[31] An anonymous reviewer raised this worry.

to other equivalences in other branches of logic. In propositional logic, disjunction is in principle definable in terms of conjunction and negation since $(A \vee B) \leftrightarrow \neg(\neg A \wedge \neg B)$ is a tautology. This fact does not entail that ‘Either London or Paris is the capital of France’ means the same as ‘It is not the case that it is not the case that London is the capital of France and it is not the case that Paris is the capital of France.’ I suggest that the same thing is true of our equivalences.[32] In conclusion, the fact that ‘Tom ought to go home’ does not have the same meaning as ‘Everyone who is perfectly rational wants Tom to go home’ is not a serious problem for the systems that include the equivalence $OA \leftrightarrow \Pi x(Rx \rightarrow \mathcal{W}_x A)$.

4.2.13. Alethic-deontic accessibility rules (ad-rules)

$T-MO$	$T-MO'$	$T-OC$	$T-OC'$
isj $\downarrow$ irj	isj jsk $\downarrow$ jrk	i $\downarrow$ isj irj where j is new	isj $\downarrow$ jsk jrk where k is new
$T-ad4$	$T-ad5$	$T-PMP$	$T-OMP$
irj jsk $\downarrow$ isk	irj isk $\downarrow$ jsk	isj irk $\downarrow$ jrl ksl where l is new	irj jsk $\downarrow$ isl lrk where l is new

Table 20

[32] In other words, meaning is stronger than necessary equivalence. The fact that A is necessarily equivalent with B does not entail that A and B have the same meaning; but if A has the same meaning as B, then A and B are necessarily equivalent.

4.2.14. Identity rules

$T-R=$	$T-S=$	$T-N=$	$T-A=$
$*$ $\downarrow$ $t=t,i$ for every t on the branch	$s=t,i$ $A[s/x],i$ $\downarrow$ $A[t/x],i$ where A is of a certain form (see below, 4.2.14)	$a=b,i$ $\downarrow$ $a=b,j$ for any j	$a=b,i$ $Aajk$ $\downarrow$ $Abjk$

Table 21[33]

$(T-S=)$ is applied only 'within worlds,' and we usually only apply the rule when A is atomic. However, we shall also allow applications of the following kind. Let M be a matrix where x_m is the first free variable in M and a_m is the constant in $M[a_1, \ldots, a, \ldots, a_n/\vec{x}]$ that replaces x_m. Furthermore, suppose we have $a=b,i$, $M[a_1, \ldots, a, \ldots, a_n/\vec{x}],i$ and $\neg Ra_m$ on the branch. Then we may apply $(T-S=)$ to obtain an extension of the branch that includes $M[a_1, \ldots, b, \ldots, a_n/\vec{x}],i$.

With the help of $(T-S=)$ and $(T-A=)$ we can prove the following theorems: $(\mathcal{W}_cB \land c=d) \rightarrow \mathcal{W}_dB$, $(\mathcal{A}_cB \land c=d) \rightarrow \mathcal{A}_dB$, $\Pi x\Pi y((\mathcal{W}_xB \land x=y) \rightarrow \mathcal{W}_yB)$ and $\Pi x\Pi y((\mathcal{A}_xB \land x=y) \rightarrow \mathcal{A}_yB)$. All of these theorems are intuitively plausible. By using $(T-N=)$, we can establish that all identities and non-identities are (absolutely and historically) necessary—i.e. we can prove all of the following theorems: $\Pi x\Pi y(x=y \rightarrow Ux=y)$, $\Pi x\Pi y(x=y \rightarrow \Box x=y)$, $\Pi x\Pi y(\neg x=y \rightarrow U\neg x=y)$, and $\Pi x\Pi y(\neg x=y \rightarrow \Box\neg x=y)$. This is plausible since every constant is treated as a rigid designator in this paper.

[33] In the identity rules R stands for 'reflexive,' S for 'substitution (of identities),' N for 'necessary identity,' and A for '(boulesic) accessibility.'

4.3 Tableau systems and some basic proof-theoretical concepts

A tableau system is a set of tableau rules. I will consider two kinds of system in this paper: (pure) boulesic systems and boulesic-deontic systems.

A (pure) (alethic) boulesic system is a tableau system that includes the propositional rules, the basic alethic rules, the basic boulesic rules, the rules for the possibilist quantifiers, the *CUTR*-rule (or *CUT*) and the identity rules. The smallest boulesic system is called $\mathcal{V}$. By adding various transfer rules, and/or boulesic, alethic and/or alethic-boulesic accessibility rules to $\mathcal{V}$, we obtain a large class of stronger boulesic systems.

A (alethic) boulesic-deontic system is a tableau system that includes $\mathcal{V}$ and all basic deontic rules. The smallest boulesic-deontic system is called $\mathcal{BD}$. Every boulesic-deontic system that includes $T-HW$, $T-\Sigma R$, $T-O\mathcal{W}$ and $T-\mathcal{W}O$ will be called a normal boulesic-deontic system. The smallest normal boulesic-deontic system is called $\mathcal{NBD}$. By adding various tableau rules from Section 4.2 to $\mathcal{BD}$, we obtain extensions of this system. Our (normal) boulesic-deontic systems illustrate how deontic logic can be 'grounded' in boulesic logic in a certain sense.

Let $aA_1, \ldots, A_n bB_1, \ldots, B_n abC_1, \ldots, C_n TrD_1, \ldots, D_n$ be the boulesic system that includes the alethic accessibility rules $A_1, \ldots, A_n$, the boulesic accessibility rules $B_1, \ldots, B_n$, the alethic-boulesic accessibility rules $C_1, \ldots, C_n$, and the transfer rules $D_1, \ldots, D_n$. A boulesic-deontic system is defined in a similar way: $aA_1, \ldots, A_n bB_1, \ldots, B_n dC_1, \ldots, C_n abD_1, \ldots,$ $D_n adE_1, \ldots, E_n TrF_1, \ldots, F_n$ is a boulesic-deontic system, where a, b, ab, and Tr are interpreted as in a boulesic system; $C_1, \ldots, C_n$ is a list (possibly empty) of deontic accessibility rules; and $E_1, \ldots, E_n$ is a list (possibly empty) of alethic-deontic rules.

Important proof theoretical concepts like the concepts of proof, theorem, derivation, consistency, inconsistency in a system, the logic of a tableau system, etc. are defined as usual (see, for example, Priest 2008).

I will now describe four different tableau systems that correspond to the four classes of models described in Definition 9. The first system is an example of a boulesic system; the other three are examples of boulesic-deontic systems.

Definition 10 *(i) Strict boulesic logic is the boulesic system that includes all a-rules and the tableau rules b4, b5, bT′, bB′, $M\mathcal{W}$, $\mathcal{W}C'$, $M\mathcal{W}'$, ab4, ab5, $\mathcal{W}MP$, and $\mathcal{A}MP$. (ii) Strong boulesic-deontic logic is the (normal) boulesic-deontic system that includes all a-rules and the tableau rules d4, d5, dT′, dB′, b4, b5, bT′, bB′, MO, OC′, MO′, ad4, ad5, OMP, PMP, $M\mathcal{W}$, $\mathcal{W}C'$, $M\mathcal{W}'$, ab4, ab5, $\mathcal{W}MP$, $\mathcal{A}MP$. (iii) Strong+ boulesic-deontic logic is the (normal) boulesic-deontic system that includes all rules in tables 12–15 and 18–20. (iv) Almost complete boulesic-deontic logic is the (normal) boulesic-deontic system that includes all rules that are contained in Strong+ boulesic-deontic logic plus $T-FTR$.*[34]

Note that the following relations hold between these systems: Strict boulesic logic $\subseteq$ Strong boulesic-deontic logic $\subseteq$ Strong+ boulesic-deontic logic $\subseteq$ Almost complete boulesic-deontic logic. As far as I can see, the following relations also hold: Strict boulesic logic $\subset$ Strong boulesic-deontic logic $\subset$ Strong+ boulesic-deontic logic $\subset$ Almost complete boulesic-deontic logic. However, I will only offer the latter claim as a conjecture in the present paper.

5. Examples of theorems

In this section, I will mention some sentences that can be proved in various systems. The informal reading of the theses should be obvious. Every formula in Table 22 is a theorem in every system in this paper; every sentence in Table 23 is a theorem in every system that includes the tableau rule $T-bD$, etc.

All of the following sentences (schemas) are theorems in every system in this paper: $\Pi x(Rx \rightarrow (\mathcal{W}_x B \leftrightarrow \neg\mathcal{A}_x\neg B))$, $\Pi x(Rx \rightarrow (\neg\mathcal{W}_x B \leftrightarrow \mathcal{A}_x\neg B))$, $\Pi x(Rx \rightarrow (\mathcal{W}_x\neg B \leftrightarrow \neg\mathcal{A}_x B))$ and $\Pi x(Rx \rightarrow (\mathcal{A}_x B \leftrightarrow \neg\mathcal{W}_x\neg B))$. Note that universal necessity is stronger than historical necessity and that universal

[34] Some of the rules in these systems are ‘redundant,’ and there are several ‘weaker’ systems that are deductively equivalent—i.e. they contain exactly the same theorems. ‘Weaker system’ here means a system with fewer primitive rules, not a system with fewer theorems.

possibility is weaker than historical possibility in every system in this paper. In other words, $UA \to \Box A$ and $\Diamond A \to MA$ are theorems in every system in this paper, while $\Box A \to UA$ and $MA \to \Diamond A$ are not theorems in any system in this paper. U and M behave as so-called $S5$-operators in every system in this paper and $\Box$ and $\Diamond$ behave as $S5$-operators in every system that includes every rule in Table 12 (note that not all rules have to be primitive).

Theorem	System
$\Pi x(Rx \to (\mathcal{W}_x(A \wedge B) \leftrightarrow (\mathcal{W}_xA \wedge \mathcal{W}_xB)))$	Every
$\Pi x(Rx \to ((\mathcal{W}_xA \vee \mathcal{W}_xB) \to \mathcal{W}_x(A \vee B)))$	Every
$\Pi x(Rx \to (\mathcal{A}_x(A \wedge B) \to (\mathcal{A}_xA \wedge \mathcal{A}_xB)))$	Every
$\Pi x(Rx \to (\mathcal{A}_x(A \vee B) \leftrightarrow (\mathcal{A}_xA \vee \mathcal{A}_xB)))$	Every
$\Pi x(Rx \to (\mathcal{W}_x(A \to B) \to (\mathcal{W}_xA \to \mathcal{W}_xB)))$	Every
$\Pi x(Rx \to (\mathcal{W}_x(A \to B) \to (\mathcal{A}_xA \to \mathcal{A}_xB)))$	Every
$\Pi x(Rx \to (\mathcal{W}_x(A \to B) \to (\mathcal{W}_x\neg B \to \mathcal{W}_x\neg A)))$	Every
$\Pi x(Rx \to (\mathcal{W}_x(A \leftrightarrow B) \to (\mathcal{W}_xA \leftrightarrow \mathcal{W}_xB)))$	Every
$\Pi x(Rx \to (\mathcal{W}_x(A \leftrightarrow B) \to (\mathcal{A}_xA \leftrightarrow \mathcal{A}_xB)))$	Every
$\Pi x(Rx \to (\mathcal{W}_x(A \leftrightarrow B) \to (\mathcal{W}_x\neg A \leftrightarrow \mathcal{W}_x\neg B)))$	Every

Table 22

Theorem	System
$\Pi x(Rx \to (\mathcal{W}_xB \to \mathcal{A}_xB))$	bD
$\Pi x(Rx \to \neg(\mathcal{W}_xB \wedge \mathcal{W}_x\neg B))$	bD
$\Pi x(Rx \to (\mathcal{A}_xB \vee \mathcal{A}_x\neg B))$	bD
$\Pi x(Rx \to \neg(\mathcal{W}_x(A \vee B) \wedge (\mathcal{W}_x\neg A \wedge \mathcal{W}_x\neg B)))$	bD
$\Pi x(Rx \to (\mathcal{W}_x(A \to B) \to (\mathcal{W}_xA \to \mathcal{A}_xB)))$	bD
$\Pi x(Rx \to (\mathcal{W}_x(A \to B) \to (\mathcal{W}_x\neg B \to \neg\mathcal{W}_xA)))$	bD

Table 23

Theorem	Systems
$\Pi x((Rx \wedge \mathcal{W}_x Rx) \rightarrow (\mathcal{W}_x B \rightarrow \mathcal{W}_x \mathcal{W}_x B))$	$b4$
$\Pi x((Rx \wedge \mathcal{W}_x Rx) \rightarrow (\mathcal{A}_x B \rightarrow \mathcal{W}_x \mathcal{A}_x B))$	$b5$
$\Pi x((Rx \wedge \mathcal{W}_x Rx) \rightarrow \mathcal{W}_x(\mathcal{W}_x B \rightarrow B))$	bT'
$\Pi x((Rx \wedge \mathcal{W}_x Rx) \rightarrow \mathcal{W}_x(\mathcal{A}_x \mathcal{W}_x A \rightarrow A))$	$bB'b4$
$\Pi x(Rx \rightarrow (\mathcal{W}_x B \rightarrow \mathcal{W}_x \mathcal{W}_x B))$	$b4UR$
$\Pi x(Rx \rightarrow (\mathcal{A}_x B \rightarrow \mathcal{W}_x \mathcal{A}_x B))$	$b5UR$
$\Pi x(Rx \rightarrow \mathcal{W}_x(\mathcal{W}_x B \rightarrow B))$	$bT'UR$
$\Pi x(Rx \rightarrow \mathcal{W}_x(\mathcal{A}_x \mathcal{W}_x A \rightarrow A))$	$bB'UR$

Table 24

Theorems	Systems
$\Pi x \Pi y((Rx \wedge Ry) \rightarrow (\mathcal{W}_x B \rightarrow \mathcal{W}_y B))$	HW
$\Pi x(Rx \rightarrow (\mathcal{W}_x B \rightarrow \Pi y(Ry \rightarrow \mathcal{W}_y B))$	HW
$\Pi x \Pi y((Rx \wedge Ry) \rightarrow (\mathcal{A}_x B \rightarrow \mathcal{A}_y B))$	HW
$\Pi x(Rx \rightarrow (\mathcal{A}_x B \rightarrow \Pi y(Ry \rightarrow \mathcal{A}_y B)))$	HW
$\Sigma x(Rx \wedge \mathcal{W}_x B) \rightarrow \Pi x(Rx \rightarrow \mathcal{W}_x B)$	HW
$\Sigma x(Rx \wedge \mathcal{A}_x B) \rightarrow \Pi x(Rx \rightarrow \mathcal{A}_x B)$	HW
$\neg \Sigma x \Sigma y((Rx \wedge Ry) \wedge (\mathcal{W}_x B \wedge \mathcal{W}_y \neg B))$	$HWbD$
$\Pi x(Rx \rightarrow (\mathcal{W}_x B \rightarrow \Pi y(Ry \rightarrow \mathcal{A}_y B)))$	$HWbD$

Table 25

Theorems	Systems
$\Pi x(Rx \rightarrow (\Box A \rightarrow \mathcal{W}_x A))$	$abM\mathcal{W}$
$\Pi x(Rx \rightarrow (\mathcal{W}_x A \rightarrow \Diamond A))$	$ab\mathcal{W}C$
$\Pi x((Rx \wedge \mathcal{W}_x Rx) \rightarrow \mathcal{W}_x(\Box A \rightarrow \mathcal{W}_x A))$	$abM\mathcal{W}'$
$\Pi x((Rx \wedge \mathcal{W}_x Rx) \rightarrow \mathcal{W}_x(\mathcal{W}_x A \rightarrow \Diamond A))$	$ab\mathcal{W}C'$
$\Pi x(Rx \rightarrow (\mathcal{W}_x A \rightarrow \Box \mathcal{W}_x A))$	$ab4UR$

$\Pi x(Rx \rightarrow (\mathcal{A}_x B \rightarrow \Box\mathcal{A}_x B))$	$ab5UR$
$\Pi x(Rx \rightarrow (\mathcal{A}_x \Box B \rightarrow \Box\mathcal{A}_x B))$	$ab\mathcal{A}MPUR$
$\Pi x(Rx \rightarrow (\mathcal{W}_x \Box A \rightarrow \Box\mathcal{W}_x A))$	$ab\mathcal{W}MPUR$
$\Pi x(Rx \rightarrow \mathcal{W}_x(\Box A \rightarrow \mathcal{W}_x A))$	$abM\mathcal{W}'UR$
$\Pi x(Rx \rightarrow \mathcal{W}_x(\mathcal{W}_x A \rightarrow \Diamond A))$	$ab\mathcal{W}C'UR$

Table 26

Name	Theorem	Systems
MO	$\Box A \rightarrow OA$	$adMO$
OC	$OA \rightarrow \Diamond A$	$adOC$
OC'	$O(OA \rightarrow \Diamond A)$	$adOC'$
MO'	$O(\Box A \rightarrow OA)$	$adMO'$
$ad4$	$OA \rightarrow \Box OA$	$ad4$
$ad5$	$PA \rightarrow \Box PA$	$ad5$
PMP	$P\Box A \rightarrow \Box PA$	$adPMP$
OMP	$O\Box A \rightarrow \Box OA$	$adOMP$

Table 27

Theorem	System
$\Pi x(Rx \rightarrow (\Pi y\mathcal{W}_x B \leftrightarrow \mathcal{W}_x \Pi yB))$	Every
$\Pi x(Rx \rightarrow (\Sigma y\mathcal{A}_x B \leftrightarrow \mathcal{A}_x \Sigma yB))$	Every
$\Pi x(Rx \rightarrow (\mathcal{A}_x \Pi yB \rightarrow \Pi y\mathcal{A}_x B))$	Every
$\Pi x(Rx \rightarrow (\Sigma y\mathcal{W}_x B \rightarrow \mathcal{W}_x \Sigma yB))$	Every

Table 28

Theorem	System
$\Pi x(Rx \rightarrow (\Box(A \rightarrow B) \rightarrow (\mathcal{W}_x A \rightarrow \mathcal{W}_x B)))$	$M\mathcal{W}$
$\Pi x(Rx \rightarrow (\Box(A \rightarrow B) \rightarrow (\mathcal{A}_x A \rightarrow \mathcal{A}_x B)))$	$M\mathcal{W}$
$\Pi x(Rx \rightarrow (\Box(A \rightarrow B) \rightarrow (\mathcal{W}_x \neg B \rightarrow \mathcal{W}_x \neg A)))$	$M\mathcal{W}$

$\Pi x(Rx \to (\Box(A \leftrightarrow B) \to (\mathcal{W}_x A \leftrightarrow \mathcal{W}_x B)))$	$M\mathcal{W}$
$\Pi x(Rx \to (\Box(A \leftrightarrow B) \to (\mathcal{A}_x A \leftrightarrow \mathcal{A}_x B)))$	$M\mathcal{W}$
$\Pi x(Rx \to (\Box(A \leftrightarrow B) \to (\mathcal{W}_x \neg A \leftrightarrow \mathcal{W}_x \neg B)))$	$M\mathcal{W}$
$\Pi x(Rx \to (\Box(A \leftrightarrow B) \to (\neg\mathcal{W}_x A \leftrightarrow \neg\mathcal{W}_x B)))$	$M\mathcal{W}$

Table 29

Theorem	System
$OA \to \Pi x(Rx \to \mathcal{W}_x A)$	$O\mathcal{W}$
$\Pi x(Rx \to \mathcal{W}_x A) \to OA$	$\mathcal{W}OHW\Sigma R$
$OA \leftrightarrow \Pi x(Rx \to \mathcal{W}_x A)$	$O\mathcal{W}\mathcal{W}OHW\Sigma R$
$PB \to \Pi x(Rx \to \mathcal{A}_x B)$	$\mathcal{W}OHW$
$\Pi x(Rx \to \mathcal{A}_x B) \to PB$	$O\mathcal{W}HW\Sigma R$
$PB \leftrightarrow \Pi x(Rx \to \mathcal{A}_x B)$	$\mathcal{W}OO\mathcal{W}HW\Sigma R$
$FA \to \Pi x(Rx \to \mathcal{W}_x \neg A)$	$O\mathcal{W}$
$\Pi x(Rx \to \mathcal{W}_x \neg A) \to FA$	$\mathcal{W}OHW\Sigma R$
$FA \leftrightarrow \Pi x(Rx \to \mathcal{W}_x \neg A)$	$O\mathcal{W}\mathcal{W}OHW\Sigma R$

Table 30

5.1. Examples: Valid arguments and valid and invalid formulas

In this section, I will consider one example of a valid argument, one example of a valid sentence and one example of an invalid sentence. I will show that *argument* 3 described in the introduction is valid (in the class of all models that satisfy $C-M\mathcal{W}$). This illustrates one of the possible applications of the systems that are introduced in this paper, namely as a tool in the analysis and evaluation of various arguments. *Argument* 3 is intuitively valid, but it seems impossible to prove this in any other systems in the literature. Nonetheless, we can prove that the conclusion is derivable from the premises in all systems in this paper that include $T-M\mathcal{W}$. Since the smallest boulesic system that includes $T-M\mathcal{W}$ is sound with respect to the class of all models that satisfy $C-M\mathcal{W}$, the argument is valid in the class of all models that satisfy $C\!-\!M\mathcal{W}$. Hence, we seem to need systems of the kind developed in this paper.

Argument 3 can be symbolised in the following way. $\Pi x(Px \to \mathcal{W}_x Mx)$ (for every x: if x is a person in the class, then x wants it to be the case that x passes the exam), Ps (Sandra is a person in the class), $\Box(Ms \to Ss)$ (it is necessary that Sandra passes the exam only if she studies hard), $Rs \to \mathcal{W}_s Ss$ (if Sandra is perfectly rational, she wants to study hard). To prove that the conclusion is derivable from the premises, we construct a semantic tableau that begins with all premises and the negation of the conclusion. Since this tableau is closed, it constitutes a derivation of the conclusion from the premises in the smallest boulesic system that includes $T\!-\!M\mathcal{W}$. Hence, the conclusion follows from the premises in the class of all models that satisfy $C\!-\!M\mathcal{W}$ (by the soundness theorems in Section 6). Here is our proof. (‘MP’ stands for the derived rule ‘Modus Ponens.’)

(1) $\Pi x(Px \to \mathcal{W}_x Mx)$, 0

(2) Ps, 0

(3) $\Box(Ms \to Ss)$, 0

(4) $\neg(Rs \to \mathcal{W}_s Ss)$, 0

(5) Rs, 0 [4, $\neg\!\to$]

(6) $\neg\mathcal{W}_s Ss$, 0 [4, $\neg\!\to$]

(7) $\mathcal{A}_s \neg Ss$, 0 [5, 6, $\neg\mathcal{W}$]

(8) $Ps \to \mathcal{W}_s Ms$, 0 [1, Π [s/x]]

(9) $\mathcal{W}_s Ms$, 0 [2, 8, MP]

(10) $0As1$ [5, 7, $\mathcal{A}$]

(11) $\neg Ss$, 1 [5, 7, $\mathcal{A}$]

(12) Ms, 1 [5, 9, 10, $\mathcal{W}$]

(13) $0r1$ [10, $T\!-\!M\mathcal{W}$]

(14) $Ms \to Ss$, 1 [3, 13, $\Box$]

(15) Ss, 1 [12, 14, MP]

(16) $*$ [11, 15]

Let us now turn to our valid sentence. In the introduction, we considered several interpretations of the so-called hypothetical imperative. One of the readings was (7): $U\Pi x(Rx \to ((\mathcal{W}_x A \land \Box(A \to B)) \to OB))$. Intuitively, this sentence says that it is absolutely necessary that if x is perfectly rational, then if x wants it to be the case that A and it is necessary that A only if B then it ought to be the case that B. Here is an instance of this schema: if x is perfectly rational and x wants to achieve end E and it is necessary that x achieves end E only if x does action A then x ought to do action A. Or more concretely, if x is perfectly rational then if x wants to become a doctor of philosophy (sometime in the future) and it is necessary that x will become a doctor of philosophy (sometime in the future) only if x writes a dissertation then x ought to write a dissertation. I will now show that (7) is a theorem in every boulesic-deontic system that includes the rules $T-\mathcal{W}O$, $T-HW$ and $T-M\mathcal{W}$. Here is our tableau proof:[35]

$$(1)\ \neg U\Pi x(Rx \to ((\mathcal{W}_x A \land \Box(A \to B)) \to OB)),\ 0$$
$$(2)\ M\neg\Pi x(Rx \to ((\mathcal{W}_x A \land \Box(A \to B)) \to OB)),\ 0\ [1,\ \neg U]$$
$$(3)\ \neg\Pi x(Rx \to ((\mathcal{W}_x A \land \Box(A \to B)) \to OB)),\ 1\ [2,\ M]$$
$$(4)\ \Sigma x\neg(Rx \to ((\mathcal{W}_x A \land \Box(A \to B)) \to OB)),\ 1\ [3,\ \neg\Pi]$$
$$(5)\ \neg(Rc \to ((\mathcal{W}_c A[c/x] \land \Box(A[c/x] \to B[c/x])) \to OB[c/x])),\ 1\ [4,\ \Sigma]$$
$$(6)\ Rc,\ 1\ [5,\ \neg\!\to]$$
$$(7)\ \neg((\mathcal{W}_c A[c/x] \land \Box(A[c/x] \to B[c/x])) \to OB[c/x]),\ 1\ [5,\ \neg\!\to]$$
$$(8)\ \mathcal{W}_c A[c/x] \land \Box(A[c/x] \to B[c/x]),\ 1\ [7,\ \neg\!\to]$$
$$(9)\ \neg OB[c/x],\ 1\ [7,\ \neg\!\to]$$
$$(10)\ \mathcal{W}_c A[c/x],\ 1\ [8,\ \land]$$
$$(11)\ \Box(A[c/x] \to B[c/x]),\ 1\ [8,\ \land]$$
$$(12)\ P\neg B[c/x],\ 1\ [9,\ \neg O]$$
$$(13)\ 1s2\ [12,\ P]$$
$$(14)\ \neg B[c/x],\ 2\ [12,\ P]$$
$$(15)\ 1Ad2\ [13,\ T-\mathcal{W}O]$$
$$(16)\ 1Ac2\ [15,\ T-HW]$$
$$(17)\ A[c/x],\ 2\ [6,\ 10,\ 16,\ \mathcal{W}]$$
$$(18)\ 1r2\ [16,\ T-M\mathcal{W}]$$
$$(19)\ A[c/x] \to B[c/x],\ 2\ [11,\ 18,\ \Box]$$

[35] In a strict sense, this is not a proof, but a proof schema. For it includes expressions such as $A[c/x]$. However, this schema shows that any proof of this form is correct.

$$(20)\ B[c/x],\ 2\ [17,\ 19,\ MP]$$
$$(21)\ *\ [14,\ 20]$$

The tableau above is closed. Hence, $U\Pi x(Rx \rightarrow ((\mathcal{W}_x A \wedge \Box(A \rightarrow B)) \rightarrow OB))$ is a theorem in every boulesic-deontic system that includes the rules $T-\mathcal{W}O$, $T-HW$ and $T-M\mathcal{W}$. It follows, by the soundness results in Section 6, that $U\Pi x(Rx \rightarrow ((\mathcal{W}_x A \wedge \Box(A \rightarrow B)) \rightarrow OB))$ is valid in the class of all models that satisfy $C-\mathcal{W}O$, $C-HW$ and $C-M\mathcal{W}$. Even though $U\Pi x(Rx \rightarrow ((\mathcal{W}_x A \wedge \Box(A \rightarrow B)) \rightarrow OB))$ is valid in some systems, $U\Pi x((\mathcal{W}_x A \wedge \Box(A \rightarrow B)) \rightarrow OB)$ is not a theorem in any system in this paper (as I mentioned in the introduction). This is as it should be since this formula has countless counterintuitive consequences. Consider, for example, the following 'instance': 'If c wants to destroy the City Hall and it is necessary that c uses a bomb to destroy the City Hall, then c ought to use a bomb to destroy the City Hall.' Suppose that the antecedent is true. Then c ought to use a bomb to destroy the City Hall. But this is absurd.

Now I will show how it is possible to use semantic tableaux to prove that a sentence is not valid and how it is possible to use open complete branches to read off countermodels. Consider the following sentence:

> For every individual x, if x wants to quench her thirst and it is necessary that x quenches her thirst only if x drinks some water, then x wants to drink some water.

This sentence can be symbolised in the following way: $\Pi x((\mathcal{W}_x Qx \wedge \Box(Qx \rightarrow Dx)) \rightarrow \mathcal{W}_x Dx)$, where Qx says that x quenches her thirst and Dx says that x drinks some water. I will show that this formula is *not* valid in the class of all models. To establish this, I will show that the formula is not a theorem in our weakest system. By the completeness theorems in Section 6, it follows that the sentence is not valid in the class of all models. I will use an open branch in a complete tree for the formula to read off a countermodel and I will verify that this model is a countermodel to the formula. In fact, it is possible to prove that the sentence is not a theorem in any system in this paper. Consequently, it is possible to show that the formula is not valid in any class of models (in this paper). It is left to the reader to verify this claim. Here is our tableau:

$$(1)\ \neg\Pi x((\mathcal{W}_x Qx \wedge \Box(Qx \to Dx)) \to \mathcal{W}_x Dx),\ 0$$
$$(2)\ \Sigma x\neg((\mathcal{W}_x Qx \wedge \Box(Qx \to Dx)) \to \mathcal{W}_x Dx),\ 0\ [1,\ \neg\Pi]$$
$$(3)\ \neg((\mathcal{W}_c Qc \wedge \Box(Qc \to Dc)) \to \mathcal{W}_c Dc),\ 0\ [2,\ \Sigma]$$
$$(4)\ \mathcal{W}_c Qc \wedge \Box(Qc \to Dc),\ 0\ [3,\ \neg\!\to]$$
$$(5)\ \neg\mathcal{W}_c Dc,\ 0\ [3,\ \neg\!\to]$$
$$(6)\ \mathcal{W}_c Qc,\ 0\ [4,\ \wedge]$$
$$(7)\ \Box(Qc \to Dc),\ 0\ [4,\ \wedge]$$
$$\swarrow \qquad\qquad \searrow$$
$$(8)\ Rc,\ 0 \qquad\qquad (9)\ \neg Rc,\ 0\ [CUTR]$$
$$(10)\ c = c,\ 0\ [T{-}R{=}]$$

It is possible to extend the left branch in this tree. Nevertheless, at this stage we cannot apply any more rules to the right branch, which is open (and complete). It follows that the whole tableau is open (and complete). Hence, $\Pi x((\mathcal{W}_x Qx \wedge \Box(Qx \to Dx)) \to \mathcal{W}_x Dx)$ is *not* a theorem in our weakest system. Consequently, the formula is *not* valid in the class of all models (by the completeness results in Section 6).

Let us verify this conclusion. We can use the right branch to read off a countermodel, $\mathcal{M}$, since this branch is open and complete. The matrix of $\mathcal{W}_c Qc$ is $\mathcal{W}_{x_1} Qx_2$ and the matrix of $\mathcal{W}_c Dc$ is $\mathcal{W}_{x_1} Dx_2$.

$W = \{\omega_0\}$, $D = \{[c]\}$, $v(c) = [c]$, and the extensions of Q and D are empty in ω_0. $\mathfrak{R}$, $\mathfrak{A}$ (and $\mathfrak{S}$) are empty. $v_{\omega_0}(\mathcal{W}_{x_1} Qx_2)$ is the extension of $\mathcal{W}_{x_1} Qx_2$ in ω_0 and $v_{\omega_0}(\mathcal{W}_{x_1} Dx_2)$ is the extension of $\mathcal{W}_{x_1} Dx_2$ in ω_0. If $\neg Ra_m, i$ is on the branch $\mathcal{B}$ and M is an n-place matrix with instantiations on the branch (where x_m is the first free variable in M and a_m is the constant in $M[a_1, \ldots, a_n/x_1, \ldots, x_n]$ that replaces x_m), then $\langle [a_1], \ldots, [a_n] \rangle$ is an element of $v_{\omega_i}(M)$ iff $M[a_1, \ldots, a_n/x_1, \ldots, x_n], i$ occurs on $\mathcal{B}$.

$\neg Rc$, 0 is on the branch, while $\mathcal{W}_{x_1} Dx_2[c,\ c/x_1,\ x_2]$, 0 ($=\mathcal{W}_c Dc$, 0) is not on the branch. x_1 is the first free variable in $\mathcal{W}_{x_1} Dx_2$ and c is the constant in $\mathcal{W}_{x_1} Dx_2[c,\ c/x_1,\ x_2]$ that replaces x_1. Consequently, $\langle [c],\ [c] \rangle$ is not an element in $v_{\omega_0}(\mathcal{W}_{x_1} Dx_2)$ ($v_{\omega_0}(\mathcal{W}_{x_1} Dx_2)$ is empty). Rc is false in ω_0 in $\mathcal{M}$, for $\neg Rc$, 0 is on $\mathcal{B}$. If Rc is false in ω_0 in $\mathcal{M}$, then $\mathcal{W}_{x_1} Dx_2[c,\ c/x_1,\ x_2]$ is true in ω_0 in $\mathcal{M}$ iff $\langle v(c),\ v(c) \rangle$ is in $v_{\omega_0}(\mathcal{W}_{x_1} Dx_2)$. Hence, $\mathcal{W}_{x_1} Dx_2[c,\ c/x_1,\ x_2]$ is true in ω_0 in $\mathcal{M}$ iff $\langle v(c),\ v(c) \rangle$ is in $v_{\omega_0}(\mathcal{W}_{x_1} Dx_2)$. $\langle v(c),\ v(c) \rangle$ is not in $v_{\omega_0}(\mathcal{W}_{x_1} Dx_2)$.

Consequently, it is not the case that $\mathcal{W}_{x_1}Dx_2[c, c/x_1, x_2]$ is true in ω_0 in $\mathcal{M}$. $\mathcal{W}_{x_1}Dx_2[c, c/x_1, x_2] = \mathcal{W}_cDc$. It follows that it is not the case that $\mathcal{W}_cDc$ is true in ω_0 in $\mathcal{M}$, that is, $\mathcal{W}_cDc$ is false in ω_0 in $\mathcal{M}$.

$\mathcal{W}_{x_1}Qx_2[c, c/x_1, x_2], 0$ (that is, $\mathcal{W}_cQc, 0$) is on the branch. x_1 is the first free variable in $\mathcal{W}_{x_1}Qx_2$ and c is the constant in $\mathcal{W}_{x_1}Qx_2[c, c/x_1, x_2]$ that replaces x_1. Hence, $\langle[c], [c]\rangle$ is an element in $v_{\omega 0}(\mathcal{W}_{x_1}Qx_2)$. If Rc is false in ω_0 in $\mathcal{M}$, then $\mathcal{W}_{x_1}Qx_2[c, c/x_1, x_2]$ is true in ω_0 in $\mathcal{M}$ iff $\langle v(c), v(c)\rangle$ is in $v_{\omega 0}(\mathcal{W}_{x_1}Qx_2)$. Accordingly, $\mathcal{W}_{x_1}Qx_2[c, c/x_1, x_2]$ is true in ω_0 in $\mathcal{M}$ iff $\langle v(c), v(c)\rangle$ is in $v_{\omega 0}(\mathcal{W}_{x_1}Qx_2)$. $\langle v(c), v(c)\rangle$ is in $v_{\omega 0}(\mathcal{W}_{x_1}Qx_2)$. Therefore, $\mathcal{W}_{x_1}Qx_2[c, c/x_1, x_2]$ is true in ω_0 in $\mathcal{M}$. $\mathcal{W}_{x_1}Qx_2[c, c/x_1, x_2] = \mathcal{W}_cQc$. Consequently, $\mathcal{W}_cQc$ is true in ω_0 in $\mathcal{M}$.

Since no possible world is alethically accessible from ω_0 in $\mathcal{M}$, $\Box(Qc \to Dc)$ is true in ω_0 in $\mathcal{M}$.

We have established that $\mathcal{W}_cQc$ is true in ω_0 in $\mathcal{M}$ and that $\Box(Qc \to Dc)$ is true in ω_0 in $\mathcal{M}$. Accordingly, $\mathcal{W}_cQc \land \Box(Qc \to Dc)$ is true in ω_0 in $\mathcal{M}$. Furthermore, we have shown that $\mathcal{W}_cDc$ is false in ω_0 in $\mathcal{M}$. It follows that $(\mathcal{W}_cQc \land \Box(Qc \to Dc)) \to \mathcal{W}_cDc$ is false in ω_0 in $\mathcal{M}$. Since $[c]$ is an object in the domain, we conclude that $\Pi x((\mathcal{W}_xQx \land \Box(Qx \to Dx)) \to \mathcal{W}_xDx)$ is false in ω_0 in $\mathcal{M}$. It follows that this formula is not valid in the class of all models. This result is intuitively plausible. If some individual is not perfectly rational, it is possible that she wants something, A, without wanting the necessary means to A. This is compatible with the proposition that several other versions of the hypothetical imperative are valid (in some models) (see above, the introduction and Section 3.6).

6. Soundness and completeness theorems

In this section, I will prove that every system in this essay is sound and complete with respect to its semantics. The concepts of soundness and completeness are defined as usual (see, for example, Priest, 2008). Many steps in the proofs are easy modifications of existing proofs. However, due to the presence of the boulesic operators in our language, some steps require some new techniques.

Lemma 11 (*Locality*): *Let $\mathcal{M}_1 = \langle D, W, \mathfrak{R}, \mathfrak{A}, \mathfrak{S}, v_1\rangle$ and $\mathcal{M}_2 = \langle D, W, \mathfrak{R}, \mathfrak{A}, \mathfrak{S}, v_2\rangle$ be two supplemented models (the lemma for unsupplemented models is similar). The language of the two, which we call $\mathcal{L}$, is the same, for they have the same domain. Let A be any closed formula of $\mathcal{L}$ such that v_1 and v_2 agree on the denotations of all the predicates, constants and matrices in it. Then for all $\omega \in W$: $v_{1\omega}(A) = v_{2\omega}(A)$.*

Proof. The proof is by recursion on the sentences in our language. 'the IH' refers to the induction hypothesis.

Atomic formulas. $v_{1\omega}(Pa_1 \ldots a_n) = 1$ iff $\langle v_1(a_1), \ldots, v_1(a_n)\rangle \in v_{1\omega}(P)$ iff $\langle v_2(a_1), \ldots, v_2(a_n)\rangle \in v_{2\omega}(P)$ iff $v_{2\omega}(Pa_1 \ldots a_n) = 1$.

Suppose that $v_{1\omega}(Ra_m) = 0$, that M is a matrix where x_m is the first free variable in M and that a_m is the constant in $M[a_1, \ldots, a_n/\vec{x}]$ that replaces x_m. Then: $v_{2\omega}(Ra_m) = 0$ and $v_{1\omega}(M[a_1, \ldots, a_n/\vec{x}] = 1$ iff $\langle v_1(a_1), \ldots, v_1(a_n)\rangle \in v_{1\omega}(M)$ iff $\langle v_2(a_1), \ldots, v_2(a_n)\rangle \in v_{2\omega}(M)$ iff $v_{2\omega}(M[a_1, \ldots, a_n/\vec{x}]) = 1$.

Truth-functional connectives. Straightforward.

($\Box$). $v_{1\omega}(\Box B) = 1$ iff for all ω' such that $\mathfrak{R}\omega\omega'$, $v_{1\omega'}(B) = 1$ iff for all ω' such that $\mathfrak{R}\omega\omega'$, $v_{2\omega'}(B) = 1$ [the IH] iff $v_{2\omega}(\Box B) = 1$.

The cases for the other alethic and deontic operators are similar.

($\mathcal{W}_cB$). A is of the form $\mathcal{W}_cB$. Suppose $v_{1\omega}(\mathcal{W}_cB) = 1$. We have two cases: $v_{1\omega}(Rc) = 0$ or $v_{1\omega}(Rc) = 1$. Suppose $v_{1\omega}(Rc) = 0$. Then $v_{2\omega}(Rc) = 0$. Hence, $v_{2\omega}(\mathcal{W}_cB) = 1$. And vice versa. Suppose $v_{1\omega}(Rc) = 1$. Then for all ω' such that $\mathfrak{A}v_1(c)\omega\omega'$: $v_{1\omega'}(B) = 1$. Accordingly, for all ω' such that $\mathfrak{A}v_2(c)\omega\omega'$: $v_{2\omega'}(B) = 1$ [by assumption and the IH]. Furthermore, $v_{2\omega}(Rc) = 1$. Hence, $v_{2\omega}(\mathcal{W}_cB) = 1$. And vice versa. Consequently, $v_{1\omega}(\mathcal{W}_cB) = 1$ iff $v_{2\omega}(\mathcal{W}_cB) = 1$.

The case for $\mathcal{A}_cB$ is similar.

(Π). $v_{1\omega}(\Pi xB) = 1$ iff for all $k_d \in \mathcal{L}(\mathcal{M})$, $v_{1\omega}(B[k_d/x]) = 1$ iff for all $k_d \in \mathcal{L}(\mathcal{M})$, $v_{2\omega}(B[k_d/x]) = 1$ [by the IH, and the fact that $v_{1\omega}(k_d) = v_{2\omega}(k_d) = d$] iff $v_{2\omega}(\Pi xB) = 1$.

The case for the particular quantifier is similar. ■

Lemma 12 (*Denotation*): *Let $\mathcal{M} = \langle D, W, \mathfrak{R}, \mathfrak{A}, \mathfrak{S}, v\rangle$ be any supplemented model (the lemma for unsupplemented models is similar). Let A be any formula of $\mathcal{L}(\mathcal{M})$ with at most one free variable, x, and a and b be any two* *constants such that* $v(a) = v(b)$*. Then for any* $\omega \in W$*:* $v_\omega(A[a/x]) = v_\omega(A[b/x])$*.*

Proof. The proof is by induction on the complexity of A.

Atomic formulas. (To illustrate, we assume that the formula has one occurrence of 'a,' distinct from each a_i.) $v_\omega(Pa_1 \ldots a \ldots a_n) = 1$ iff $\langle v(a_1), \ldots, v(a), \ldots, v(a_n)\rangle \in v_\omega(P)$ iff $\langle v(a_1), \ldots, v(b), \ldots, v(a_n)\rangle \in v_\omega(P)$ iff $v_\omega(Pa_1 \ldots b \ldots a_n) = 1$.

Suppose $v_\omega(Ra_m) = 0$, that M is a matrix where x_m is the first free variable in M and that a_m is the constant in $M[a_1, \ldots, a, \ldots, a_n/\vec{x}]$ ($M[a_1, \ldots, b, \ldots, a_n/\vec{x}]$) that replaces x_m. (To illustrate, we assume that the formula has one occurrence of 'a' distinct from each a_i and that a_m is not a (b).) Then: $v_\omega(M[a_1, \ldots, a, \ldots, a_n/\vec{x}]) = 1$ iff $\langle v(a_1), \ldots, v(a), \ldots, v(a_n)\rangle \in v_\omega(M)$ iff $\langle v(a_1), \ldots, v(b), \ldots, v(a_n)\rangle \in v_\omega(M)$ iff $v_\omega(M[a_1, \ldots, b, \ldots, a_n/\vec{x}]) = 1$.

Truth-functional connectives. Straightforward.

($\Box$). $v_\omega(\Box B[a/x]) = 1$ iff for all ω' such that $\mathfrak{R}\omega\omega'$, $v_{\omega'}(B[a/x]) = 1$ iff for all ω' such that $\mathfrak{R}\omega\omega'$, $v_{\omega'}(B[b/x]) = 1$ [the IH] iff $v_\omega(\Box B[b/x]) = 1$.

The arguments for the other primitive alethic and deontic operators are similar.

($\mathcal{W}_t$). A is of the form $\mathcal{W}_t B$. Either $v_\omega(Rt) = 1$ or $v_\omega(Rt) = 0$. We have already shown that the result holds if $v_\omega(Rt) = 0$. Accordingly, suppose that $v_\omega(Rt) = 1$. Since x is the only free variable, t cannot be a variable distinct from x. So, t is either x or a constant. Suppose t is x. Then $v_\omega(\mathcal{W}_x B[a/x]) = 1$ iff $v_\omega(\mathcal{W}_a B[a/x]) = 1$ iff for all ω' such that $\mathfrak{A}v(a)\omega\omega'$, $v_{\omega'}(B[a/x]) = 1$ iff for all ω' such that $\mathfrak{A}v(b)\omega\omega'$, $v_{\omega'}(B[b/x]) = 1$ [by the fact that $v(a) = v(b)$ and the IH] iff $v_\omega(\mathcal{W}_b B[b/x]) = 1$ iff $v_\omega(\mathcal{W}_x B[b/x]) = 1$. Suppose t is a constant, say c. Then $v_\omega(\mathcal{W}_c B[a/x]) = 1$ iff for all ω' such that $\mathfrak{A}v(c)\omega\omega'$, $v_{\omega'}(B[a/x]) = 1$ iff for all ω' such that $\mathfrak{A}v(c)\omega\omega'$, $v_{\omega'}(B[b/x]) = 1$ [by the IH] iff $v_\omega(\mathcal{W}_c B[b/x]) = 1$.

The case for $\mathcal{A}_t$ is similar.

(Π). Let A be of the form $\Pi y B$. If $x = y$, then $A[a/x] = A[b/x] = A$, so the result is trivial. Accordingly, suppose that x and y are distinct. Then, $(\Pi y B)[b/x] = \Pi y(B[b/x])$ and $(B[b/x])[a/y] = (B[a/y])[b/x]$. $v_\omega((\Pi y B)[a/x])$

$= 1$ iff $v_\omega(\Pi y(B[a/x])) = 1$ iff for all $k_d \in \mathcal{L}(\mathcal{M})$, $v_\omega((B[a/x])[k_d/y]) = 1$ iff for all $k_d \in \mathcal{L}(\mathcal{M})$, $v_\omega((B[k_d/y])[a/x]) = 1$ iff for all $k_d \in \mathcal{L}(\mathcal{M})$, $v_\omega((B[k_d/y])[b/x]) = 1$ [the IH] iff for all $k_d \in \mathcal{L}(\mathcal{M})$, $v_\omega((B[b/x])[k_d/y]) = 1$ iff $v_\omega(\Pi y(B[b/x])) = 1$ iff $v_\omega((\Pi yB)[b/x]) = 1$.

The case for the particular quantifier (Σ) is similar. ■

6.1. Soundness theorem

Let $\mathcal{M} = \langle D, W, \mathfrak{R}, \mathfrak{A}, \mathfrak{S}, v\rangle$ be any (supplemented) model and $\mathcal{B}$ any branch of a tableau. Then $\mathcal{B}$ is satisfiable in $\mathcal{M}$ iff there is a function f from 0, 1, 2, . . . to W such that

(i) A is true in $f(i)$ in $\mathcal{M}$, for every node A,i on $\mathcal{B}$,

(ii) if irj is on $\mathcal{B}$, then $\mathfrak{R}f(i)f(j)$ in $\mathcal{M}$,

(iii) if isj is on $\mathcal{B}$, then $\mathfrak{S}f(i)f(j)$ in $\mathcal{M}$, and

(iv) if $iAcj$ is on $\mathcal{B}$, then $\mathfrak{A}v(c)f(i)f(j)$ in $\mathcal{M}$.

If these conditions are fulfilled, we say that f shows that $\mathcal{B}$ is satisfiable in $\mathcal{M}$.

Lemma 13 (*Soundness Lemma*): *Let $\mathcal{B}$ be any branch of a tableau and $\mathcal{M}$ be any model. If $\mathcal{B}$ is satisfiable in $\mathcal{M}$ and a tableau rule is applied to it, then there is a model $\mathcal{M}'$ and an extension of $\mathcal{B}$, $\mathcal{B}'$, such that $\mathcal{B}'$ is satisfiable in $\mathcal{M}'$.*

Proof. The proof is by induction on the height of the derivation. Let f be a function that shows that the branch $\mathcal{B}$ is satisfiable in $\mathcal{M}$.

Connectives and the modal operators. Straightforward.

($\mathcal{W}$). Suppose that Rc,i, $\mathcal{W}_cB,i$, and $iAcj$ are on $\mathcal{B}$, and that we apply the $\mathcal{W}$-rule. Then we get an extension of $\mathcal{B}$ that includes B,j. Since $\mathcal{B}$ is satisfiable in $\mathcal{M}$, $\mathcal{W}_cB$ is true in $f(i)$ and Rc is true in $f(i)$. Moreover, for any i and j such that $iAcj$ is on $\mathcal{B}$, $\mathfrak{A}v(c)f(i)f(j)$. Hence by the truth conditions for $\mathcal{W}_cB$, B is true in $f(j)$.

($\mathcal{A}$). Suppose that Rc,i, $\mathcal{A}_cB,i$ are on $\mathcal{B}$ and that we apply the $\mathcal{A}$-rule to get an extension of $\mathcal{B}$ that includes nodes of the form $iAcj$ and B,j. Since $\mathcal{B}$ is satisfiable in $\mathcal{M}$, $\mathcal{A}_cB$ is true in $f(i)$ and Rc is true in $f(i)$. Hence, for some w in W, $\mathfrak{A}v(c)f(i)w$ and B is true in w [by the truth conditions for

$\mathcal{A}_c B$ and the fact that Rc is true in $f(i)$]. Let f' be the same as f except that $f'(j) = w$. Since f and f' differ only at j, f' shows that $\mathcal{B}$ is satisfiable in $\mathcal{M}$. Moreover, by definition $\mathfrak{A}v(c)f'(i)f'(j)$, and B is true in $f'(j)$.

($\neg\mathcal{W}$) and ($\neg\mathcal{A}$). Similar.

(Π). Suppose that $\Pi xA{,}i$ is on $\mathcal{B}$ and that we apply the Π-rule to get an extension of $\mathcal{B}$ that includes a node of the form $A[a/x]{,}i$. $\mathcal{M}$ makes ΠxA true in $f(i)$. For $\mathcal{B}$ is satisfiable in $\mathcal{M}$. Hence, $A[k_d/x]$ is true in $f(i)$ in $\mathcal{M}$, for all $k_d \in \mathcal{L}(\mathcal{M})$. Let d be such that $v(a) = v(k_d)$. By the Denotation Lemma, $A[a/x]$ is true in $f(i)$ in $\mathcal{M}$. Accordingly, we can take $\mathcal{M}'$ to be $\mathcal{M}$.

(Σ). Suppose that $\Sigma xA{,}i$ is on $\mathcal{B}$ and that we apply the Σ-rule to get an extension of $\mathcal{B}$ that includes a node of the form $A[c/x]{,}i$ (where c is new). Since $\mathcal{B}$ is satisfiable in $\mathcal{M}$, ΣxA is true in $f(i)$ in $\mathcal{M}$. Hence, there is some $k_d \in \mathcal{L}(\mathcal{M})$ such that $\mathcal{M}$ makes $A[k_d/x]$ true in $f(i)$. Let $\mathcal{M}' = \langle D, W, \mathfrak{R}, \mathfrak{A}, \mathfrak{S}, v' \rangle$ be the same as $\mathcal{M}$ except that $v'(c) = d$. Since c does not occur in $A[k_d/x]$, $A[k_d/x]$ is true in $f(i)$ in $\mathcal{M}'$, by the Locality Lemma. By the Denotation Lemma and the fact that $v'(c) = d = v'(k_d)$, $A[c/x]$ is true in $f(i)$ in $\mathcal{M}'$. Furthermore, $\mathcal{M}'$ makes all other formulas on the branch true at their respective worlds as well, by the Locality Lemma. For c does not occur in any other formula on the branch.

($\neg\Pi$) and ($\neg\Sigma$). Straightforward.

Accessibility rules. I will go through three examples to illustrate the method.

($T{-}M\mathcal{W}$). Suppose we have $iAcj$ on $\mathcal{B}$, and that we apply ($T{-}M\mathcal{W}$) to obtain an extension of $\mathcal{B}$ that includes irj. Since $\mathcal{B}$ is satisfiable in $\mathcal{M}$, $\mathfrak{A}v(c)f(i)f(j)$. It follows that $\mathfrak{R}f(i)f(j)$, since $\mathcal{M}$ satisfies the condition $C{-}M\mathcal{W}$.

($T{-}\mathcal{W}C$). Suppose that i is on $\mathcal{B}$, and that we apply ($T{-}\mathcal{W}C$) to give an extended branch containing $iAcj$ and irj, where j is new. Since $\mathcal{B}$ is satisfiable in $\mathcal{M}$, $f(i)$ is in W and $v(c)$ is in D. Hence, for some w in W, $\mathfrak{A}v(c)f(i)w$ and $\mathfrak{R}f(i)w$, since $\mathcal{M}$ satisfies condition $C{-}\mathcal{W}C$. Let f' be the same as f except that $f'(j) = w$. Since j does not occur on $\mathcal{B}$, f' shows that $\mathcal{B}$ is satisfiable in $\mathcal{M}$. Moreover, $\mathfrak{A}v(c)f'(i)f'(j)$ and $\mathfrak{R}f'(i)f'(j)$ by construction. Hence, f' shows that the extension of $\mathcal{B}$ is satisfiable in $\mathcal{M}$.

($T{-}A{=}$). Suppose we have $a = b{,}i$ and $jAak$ on a branch and that we apply ($T{-}A{=}$) to obtain an extension that includes $jAbk$. Since f shows that the branch is satisfiable in $\mathcal{M}$, $a = b$ is true in $f(i)$ and $\mathfrak{A}v(a)f(j)f(k)$ in $\mathcal{M}$. Accordingly, $v(a) = v(b)$. Hence, $\mathfrak{A}v(b)f(j)f(k)$, and we may take $\mathcal{M}'$ to be $\mathcal{M}$. ■

Theorem 14 (*Soundness Theorem*): *Every system S in this paper is sound with respect to its semantics.*

Proof. Suppose that B does not follow from Γ in $\mathbf{M}$, where $\mathbf{M}$ is the class of models that corresponds to S. Then every premise in Γ is true and the conclusion B false at some world w in some model in $\mathbf{M}$. Consider an S-tableau whose initial list consists of A,0 for every $A \in \Gamma$ and $\neg B$,0, where '0' refers to w. Then the initial list is satisfiable in $\mathbf{M}$. Every time we apply a rule to this list it produces at least one extension that is satisfiable in $\mathbf{M}$ (by the Soundness Lemma). Hence, we can find a whole branch such that every initial section of this branch is satisfiable in $\mathbf{M}$. This branch cannot be closed, for then some sentence would be both true and false in some possible world in some model in $\mathbf{M}$. Accordingly, the tableau is open. Consequently, B is not derivable from Γ in S. In conclusion, if B is derivable from Γ in S, then B follows from Γ in $\mathbf{M}$. ■

6.2. Completeness theorem

In this section, I will show that every system in this paper is complete with respect to its semantics. However, first we must define some important concepts.

Informally, a complete tableau is a tableau where every rule that can be applied has been applied. Since different systems include different rules, a tableau can be complete in one system even though it is not complete in another system. Furthermore, since the tableau rules may be applied in different orders, there may be several different (complete) tableaux for the same sentence or set of sentences in one and the same system, some longer than others, some shorter. To produce a complete tableau (in our completeness proofs) we shall use the following method.[36] (1) For every open branch

[36] Note that it is often possible to produce shorter proofs by using some more intuitive method instead.

on the tree, we shall do the following. We shall begin at its root and move towards its tip. We shall apply any rule that produces something that has not appeared on the branch before. For example, there is no point in applying Σ more than once to a node of the form ΣxA,i. We shall not apply any rules to a branch that is already closed. Some rules can be applied more than once, for example, $\Box$ and Π. When we arrive at a node of the form $\Box A$,i and it is possible to apply $\Box$ several times, then we shall make all applications at once and we shall do the same for all similar nodes. (2) When we have extended all open branches on the tree in this way, we shall repeat the procedure. Some rules introduce new 'possible worlds,' for example $T-aD$ and $T-\mathcal{W}C$. If a rule introduces a new possible world, then we shall apply it once at the tip of every open branch at the end of every cycle (that is, when we have gone through all nodes). If a system includes several different rules that introduce new possible worlds ($R1$, $R2, \ldots$), we shall alternate between them. The first time, we shall use $R1$ once; the second time we shall use $R2$ once, etc. Before we conclude a cycle and start to move through all nodes again, we shall apply $CUTR$. We shall split the end of every open branch in the tree and add Rc,i to the left node and $\neg Rc$,i to the right node, for every constant c that occurs as an index to some boulesic operator on the tree and i on the branch. The tableau is *incomplete* precisely when there is still something to do according to this method. A tableau is *complete* iff it is not incomplete.

Definition 15 (*Induced Model*): *Let $\mathcal{B}$ be an open complete branch of a tableau, let i, j, k, etc. be numbers on $\mathcal{B}$, and let I be the set of numbers on $\mathcal{B}$. Furthermore, let C be the set of all constants on $\mathcal{B}$. Define $a \sim b$ to mean that $a = b$,0 is on the branch. $a \sim b$ is obviously an equivalence relation. Let $[a]$ be the equivalence class of a under $\sim$. The (supplemented) model, $\mathcal{M} = \langle D, W, \mathfrak{R}, \mathfrak{S}, \mathfrak{A}, v\rangle$, induced by $\mathcal{B}$ is defined as follows. $D = \{[a]\colon a \in C\}$ (or, if $C = \emptyset$, $D = \{o\}$ for an arbitrary o). (o is not in the extension of anything.) $W = \{\omega_i\colon$ i occurs on $\mathcal{B}\}$, $\mathfrak{R}\omega_i\omega_j$ iff irj occurs on $\mathcal{B}$, $\mathfrak{S}\omega_i\omega_j$ iff isj occurs on $\mathcal{B}$, $\mathfrak{A}v(a)\omega_i\omega_j$ iff iAaj occurs on $\mathcal{B}$. $v(a) = [a]$, and $\langle[a_1], \ldots, [a_n]\rangle \in v_{\omega_i}(P)$ iff $Pa_1 \ldots a_n$,i is on $\mathcal{B}$, given that P is any n-place predicate other than identity. If $\neg Ra_m$,i occurs on $\mathcal{B}$ and M is an n-place matrix with* *instantiations on the branch (where x_m is the first free variable in M and a_m is the constant in $M[a_1, \ldots, a_n/\vec{x}]$ that replaces x_m), then $\langle [a_1], \ldots, [a_n] \rangle \in v_{\omega_i}(M)$ iff $M[a_1, \ldots, a_n/\vec{x}]$,$i$ occurs on $\mathcal{B}$. (Due to the identity rules this is well defined.) When we have $a = b$,0, $b = c$,0, etc. we choose one single object for all constants to denote.*

Lemma 16 (*Completeness Lemma*): *Let $\mathcal{B}$ be an open branch in a complete tableau and let $\mathcal{M}$ be a (supplemented) model induced by $\mathcal{B}$. Then, for every formula A:*
(i) if A,i is on $\mathcal{B}$, then $v_{\omega_i}(A) = 1$, and
(ii) if $\neg A$,i is on $\mathcal{B}$, then $v_{\omega_i}(A) = 0$.

Proof. The proof is by induction on the complexity of A.

Atomic formulas.

$Pa_1 \ldots a_n$,i is on $\mathcal{B} \Rightarrow \langle [a_1], \ldots, [a_n] \rangle \in v_{\omega_i}(P) \Rightarrow \langle v(a_1), \ldots, v(a_n) \rangle \in v_{\omega_i}(P) \Rightarrow v_{\omega_i}(Pa_1 \ldots a_n) = 1$.

$\neg Pa_1 \ldots a_n$,i is on $\mathcal{B} \Rightarrow Pa_1 \ldots a_n$,$i$ is not on $\mathcal{B}$ ($\mathcal{B}$ open) $\Rightarrow \langle [a_1], \ldots, [a_n] \rangle \notin v_{\omega_i}(P) \Rightarrow \langle v(a_1), \ldots, v(a_n) \rangle \notin v_{\omega_i}(P) \Rightarrow v_{\omega_i}(Pa_1 \ldots a_n) = 0$.

$a = b$,i is on $\mathcal{B} \Rightarrow a \sim b$ ($T-N=$) $\Rightarrow [a] = [b] \Rightarrow v(a) = v(b) \Rightarrow v_{\omega_i}(a = b) = 1$.

$\neg a = b$,i is on $\mathcal{B} \Rightarrow a = b$,0 is not on $\mathcal{B}$ ($\mathcal{B}$ open) $\Rightarrow$ it is not the case that $a \sim b \Rightarrow [a] \neq [b] \Rightarrow v(a) \neq v(b) \Rightarrow v_{\omega_i}(a = b) = 0$.

Suppose that M is a matrix where x_m is the first free variable and a_m is the constant in $M[a_1, \ldots, a_n/\vec{x}]$ that replaces x_m and that $v_{\omega_i}(Ra_m) = 0$. Then: $M[a_1, \ldots, a_n/\vec{x}]$,$i$ occurs on $\mathcal{B} \Rightarrow \langle [a_1], \ldots, [a_n] \rangle \in v_{\omega_i}(M) \Rightarrow \langle v(a_1), \ldots, v(a_n) \rangle \in v_{\omega_i}(M) \Rightarrow v_{\omega_i}(M[a_1, \ldots, a_n/\vec{x}]) = 1$.

$\neg M[a_1, \ldots, a_n/\vec{x}]$,$i$ occurs on $\mathcal{B} \Rightarrow M[a_1, \ldots, a_n/\vec{x}]$,$i$ is not on $\mathcal{B}$ ($\mathcal{B}$ open) $\Rightarrow \langle [a_1], \ldots, [a_n] \rangle \notin v_{\omega_i}(M) \Rightarrow \langle v(a_1), \ldots, v(a_n) \rangle \notin v_{\omega_i}(M) \Rightarrow v_{\omega_i}(M[a_1, \ldots, a_n/\vec{x}]) = 0$.

Other truth-functional connectives and modal operators. Straightforward.

Boulesic operators. ($\mathcal{A}$). Suppose $\mathcal{A}_c B$,i is on $\mathcal{B}$. Furthermore, suppose that Rc,i is not on $\mathcal{B}$. Then $\neg Rc$,i is on $\mathcal{B}$ [by *CUTR* (or *CUT*)]. Hence, $\mathcal{A}_c B$ is true in ω_i by definition and previous steps. Suppose Rc,i is on $\mathcal{B}$. Then the $\mathcal{A}$-rule has been applied to $\mathcal{A}_c B$,i, since the branch is complete. So, for some new j, $iAcj$ and B,j occur on $\mathcal{B}$. By the induction hypothesis,

$\mathfrak{A}v(c)\omega_i\omega_j$, and B is true in ω_j. Since Rc,i is on $\mathcal{B}$, $v(c)$ is perfectly rational in ω_i. Hence, $\mathcal{A}_cB$ is true in ω_i, as required. Suppose $\neg\mathcal{A}_cB,i$ is on $\mathcal{B}$. Furthermore, suppose that Rc,i is not on $\mathcal{B}$. Then $\neg Rc,i$ is on $\mathcal{B}$ [by *CUTR* (or *CUT*)]. Consequently, $\mathcal{A}_cB$ is false in ω_i by definition and previous steps. Suppose Rc,i is on $\mathcal{B}$. Then the $\neg\mathcal{A}$-rule has been applied, and $\mathcal{W}_c\neg B$, i is on $\mathcal{B}$ since the branch is complete. Again, since Rc,i is on $\mathcal{B}$ and the branch is complete, the $\mathcal{W}$-rule has been applied and for every j such that $iAcj$ is on $\mathcal{B}$, $\neg B,j$ is on $\mathcal{B}$. By the induction hypothesis, B is false in every ω_j such that $\mathfrak{A}v(c)\omega_i\omega_j$. Since Rc,i is on $\mathcal{B}$, $v(c)$ is perfectly rational in ω_i. It follows that $\mathcal{A}_cB$ is false in ω_i, as required.

($\mathcal{W}$). Similar as for ($\mathcal{A}$).

Quantifiers. (Σ). Suppose that $\Sigma xA,i$ is on the branch. Since the tableau is complete (Σ) has been applied. Accordingly, for some c, $A[c/x],i$ is on the branch. Hence, $v_{\omega_i}(A[c/x]) = 1$, by (IH). For some $k_d \in \mathcal{L}(\mathcal{M})$, $v(c) = d$, and $v(k_d) = d$. Consequently, $v_{\omega_i}(A[k_d/x]) = 1$, by the Denotation Lemma. It follows that $v_{\omega_i}(\Sigma xA) = 1$. Suppose that $\neg\Sigma xA,i$ is on the branch. Since the tableau is complete ($\neg\Sigma$) has been applied. So, $\Pi x\neg A,i$ is on the branch. Again, since the tableau is complete (Π) has been applied. Thus, for all $c \in C$, $\neg A[c/x],i$ is on the branch. Consequently, $v_{\omega_i}(A[c/x]) = 0$ for all $c \in C$ [by the induction hypothesis]. If $k_d \in \mathcal{L}(\mathcal{M})$, then for some $c \in C$, $v(c) = v(k_d)$. By the Denotation Lemma, for all $k_d \in \mathcal{L}(\mathcal{M})$, $v_{\omega_i}(A[k_d/x]) = 0$. Consequently, $v_{\omega_i}(\Sigma xA) = 0$.

The case for Π is similar. ■

Theorem 17 (*Completeness Theorem*): *Every system in this paper is complete with respect to its semantics.*

Proof. First we prove that the theorem holds for our weakest system $\mathcal{V}$. Then we extend the theorem to all extensions of this system. Let **M** be the class of models that corresponds to $\mathcal{V}$.

Suppose that B is not derivable from Γ in $\mathcal{V}$: then it is not the case that there is a closed $\mathcal{V}$-tableau whose initial list comprises A,0 for every A in Γ and $\neg B$,0. Let t be a complete $\mathcal{V}$-tableau whose initial list comprises A,0 for every A in Γ and $\neg B$,0. Then t is not closed—i.e. it is open. Since t is open, there is at least one open branch in t. Let $\mathcal{B}$ be an open branch in t. The model induced by $\mathcal{B}$ makes all the premises in Γ true and B false in ω_0. Hence, it is not the case that B follows from Γ in **M**. Consequently, if B follows from Γ in $\mathcal{M}$, then B is derivable from Γ in $\mathcal{V}$.

To prove that all extensions of $\mathcal{V}$ are complete with respect to their semantics, we have to check that the model induced by the open branch $\mathcal{B}$ is of the right kind. To do this we first check that this is true for every single semantic condition. Then we combine each of the individual arguments. I will go through some steps to illustrate the method.

$C-bD$. Suppose that ω_i is in W. Then i occurs on $\mathcal{B}$ [by the definition of an induced model]. Since $\mathcal{B}$ is complete $(T-bD)$ has been applied. Hence, for some j, $iAcj$ is on $\mathcal{B}$. Accordingly, for some ω_j, $\mathfrak{A}v(c)\omega_i\omega_j$, as required [by the definition of an induced model].

$C-b4$. Suppose that $\mathfrak{A}v(c)\omega_i\omega_j$ and $\mathfrak{A}v(c)\omega_j\omega_k$. Then $iAcj$ and $jAck$ occur on $\mathcal{B}$ [by the definition of an induced model]. Since $\mathcal{B}$ is complete, $(T-b4)$ has been applied and $iAck$ occurs on $\mathcal{B}$. It follows that $\mathfrak{A}v(c)\omega_i\omega_k$, as required [by the definition of an induced model].

$C-HW$. Suppose that $\mathfrak{A}v(c)\omega_i\omega_j$. Then $iAcj$ occurs on $\mathcal{B}$ [by the definition of an induced model]. Since $\mathcal{B}$ is complete, $(T-HW)$ has been applied and $iAdj$ occurs on $\mathcal{B}$. Consequently, $\mathfrak{A}v(d)\omega_i\omega_j$, as required [by the definition of an induced model].

$C-M\mathcal{W}$. Suppose that $\mathfrak{A}v(c)\omega_i\omega_j$. Then $iAcj$ occurs on $\mathcal{B}$ [by the definition of an induced model]. Since $\mathcal{B}$ is complete, $(T-M\mathcal{W})$ has been applied and irj occurs on $\mathcal{B}$. Consequently, $\mathfrak{R}\omega_i\omega_j$, as required [by the definition of an induced model].

$C-\mathcal{W}C$. Suppose that ω_i is in W. Then i occurs on $\mathcal{B}$ [by the definition of an induced model]. Since $\mathcal{B}$ is complete $(T-\mathcal{W}C)$ has been applied. Accordingly, for some j, $iAcj$ and irj are on $\mathcal{B}$. Thus, for some ω_j, $\mathfrak{A}v(c)\omega_i\omega_j$ and $\mathfrak{R}\omega_i\omega_j$, as required [by the definition of an induced model]. ■

7. Conclusion

In this paper, I have developed a set of boulesic and boulesic-deontic tableau systems and I have investigated some possible connections between boulesic logic and deontic logic. Boulesic logic is a new kind of logic that deals with 'boulesic' concepts and expressions, such as wanting and accepting, and 'boulesic' sentences, arguments and systems. I have shown how deontic logic, the logic of norms, might be grounded in boulesic logic. I have used a kind of possible world models to define the systems semantically and I have shown that all systems are sound and complete with respect to their semantics. Intuitively, we can think of our semantics as a description of the structure of a perfectly rational will. Finally, I have mentioned some interesting theorems that can be proved in our systems, including some versions of the so-called *hypothetical imperative*.

The deontic fragments of the systems in this paper are pretty standard monadic deontic systems. For a long time, systems of this kind have been criticised and various deontic 'paradoxes' have been introduced, for example, Ross's paradox, the paradox of derived obligations, the contrary-to-duty paradox, the good Samaritan paradox, the paradox of epistemic obligation and the free choice permission paradox.[37] Some think that these puzzles show that normal deontic logic is seriously defective. However, I am inclined to believe that most of the so-called 'deontic paradoxes' can be 'solved' and that they do not show that we have to abandon classical deontic logic. Of course, some of the puzzles are quite serious, for example, the contrary-to-duty paradox. It does not seem to be possible to solve this puzzle adequately in normal monadic deontic systems. This does not necessarily imply that we have to abandon classical deontic logic, but it indicates that the systems in this paper should be expanded or supplemented.[38]

I would now like to mention two ways in which the systems in this paper can be improved.

[37] For more on deontic paradoxes, see, for example, (Åqvist 1967; Castañeda 1981; Chisholm 1963; Hilpinen and McNamara 2013; Prior 1954, 1958; Ross 1941, 1944; and von Wright 1968).

[38] In (Rönnedal 2018), I discuss the contrary-to-duty paradox and suggest a solution. This solution, which is attractive in many respects, does not require that we abandon normal monadic deontic logic. The systems in the present paper are compatible with this solution. For more on the contrary-to-duty paradox and various possible solutions, see Rönnedal (forthcoming).

First, the systems in this paper can be combined with temporal logic. In a quantified temporal alethic boulesic deontic system, it is possible to investigate ‘diachronistic’ rationality and the relationships between temporal, alethic, boulesic and deontic concepts. I am currently trying to develop a set of quantified temporal alethic boulesic deontic systems.

Second, there appears to be a close connection between the logic of wishing/not accepting and the logic of good/bad. Good and bad are usually strongly connected in formal systems to the logic of preference (see, for example, Chisholm and Sosa 1966; Lenzen 1983; and Hansson 1990). In future work, I hope that I will be able to combine boulesic logic with the logic of preference and construct a set of boulesic-preference systems. Such systems might be used to overcome some of the shortcomings with the kind of monadic systems that I have investigated in this paper. Systems of this kind might, for example, perhaps be used to solve the contrary-to-duty paradox.

No doubt there are other possible extensions, but these examples seem to me to be among the most interesting ones. I hope to return to these topics in future work.

Acknowledgements

The first sketch of what has turned into the present paper was written in 2010. However, I only seriously started to work on the project in 2016. The first version of the text was finished in 2017. I would like to thank all colleagues who have commented on earlier versions of this paper since then. I would also like to thank two anonymous reviewers for *Organon F* for some very detailed and interesting comments.

References

Åqvist, Lennart. 1967. “Good Samaritans, Contrary-to-duty Imperatives, and Epistemic Obligations.” *Noûs* 1 (4): 361–79. https://doi.org/10.2307/2214624

Åqvist, Lennart. 1987. *Introduction to Deontic Logic and the Theory of Normative Systems.* Naples: Bibliopolis.

Åqvist, Lennart. 2002. “Deontic Logic.” In *Handbook of Philosophical Logic*, 2nd Edition, vol. 8, edited by D. M. Gabbay and F. Guenthner, 147–264. Dordrecht/ Boston/London: Kluwer Academic Publishers. https://doi.org/10.1007/978-94-010-0387-2_3

Åqvist, Lennart, and Hoepelman, Jaap. 1981. “Some Theorems about a ‘Tree’ System of Deontic Tense Logic.” In *New Studies in Deontic Logic: Norms, Actions, and the Foundation of Ethics*, edited by R. Hilpinen, 187–221. Dordrecht: D. Reidel Publishing Company. https://doi.org/10.1007/978-94-009-8484-4_9

Aristotle. 1992. *The Nicomachean Ethics.* Translated by Sir David Ross. Oxford and New York: Oxford University Press.

Barcan (Marcus), Ruth. C. 1946. “A Functional Calculus of First Order Based on Strict Implication.” *Journal of Symbolic Logic* 11 (1): 1–16. https://doi.org/10.2307/2269159

Bedke, Matthew S. 2009. “The Iffiest Oughts: A Guise of Reasons Account of End-Given Conditionals.” *Ethics* 119 (4): 672–98. https://doi.org/10.1086/600130

Björklund, Fredrik, Björnsson, Gunnar, Eriksson, John, Francén Olinder, Ragnar, and Strandberg, Caj. 2012. “Recent Work on Motivational Internalism.” *Analysis* 72 (1): 124–37. https://doi.org/10.1093/analys/anr118

Björnsson, Gunnar, Strandberg, Caj, Francén Olinder, Ragnar, Eriksson, John, and Björklund, Fredrik. Eds. 2015. *Motivational Internalism.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199367955.001.0001

Blackburn, Patrick, de Rijke, Maarten, and Venema, Yde. 2001. *Modal Logic.* Cambridge University Press. https://doi.org/10.1017/CBO9781107050884

Blackburn, Patrick, van Benthem, Johan, and Wolter, Frank. Eds. 2007. *Handbook of Modal Logic.* Elsevier.

Bratman, Michael E. 1999. *Intention, Plans, and Practical Reason.* CSLI Publications.

Broersen, Jan M. 2011. “Making a Start with the *stit* Logic Analysis of Intentional Action.” *Journal of Philosophical Logic* 40 (4): 499–530. https://doi.org/10.1007/s10992-011-9190-6

Broersen, Jan M., Dastani, Mehdi, and van der Torre, Leendert. 2001. “Resolving Conflicts between Beliefs, Obligations, Intentions, and Desires.” In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, edited by Salem Benferhat, and Philippe Besnard, 568–79. Springer. https://doi.org/10.1007/3-540-44652-4_50

Broome, John. 1999. “Normative Requirements.” *Ratio* (new series) 12 (4): 398–419. https://doi.org/10.1111/1467-9329.00101

Broome, John. 2013. *Rationality through Reasoning.* Wiley-Blackwell. https://doi.org/10.1002/9781118609088

Brunero, John. 2010. “Self-Governance, Means-Ends Coherence, and Unalterable Ends.” *Ethics* 120 (3): 579–91. https://doi.org/10.1086/652448

Carnap, Rudolf. 1946. “Modalities and Quantification.” *Journal of Symbolic Logic* 11 (2): 33–64. https://doi.org/10.2307/2268610

Castañeda, Héctor-Neri. 1981. "The Paradoxes of Deontic Logic: the Simplest Solution to all of them in one Fell Swoop." In *New Studies in Deontic Logic: Norms, Actions, and the Foundation of Ethics*, edited by R. Hilpinen, 37–85. Dordrecht: D. Reidel Publishing Company. https://doi.org/10.1007/978-94-009-8484-4_2

Chellas, Brian. F. 1969. *The Logical Form of Imperatives*. Stanford: Perry Lane Press.

Chellas, Brian. F. 1980. *Modal Logic: An Introduction*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511621192

Chisholm, Roderick M. 1963. "Contrary-to-duty Imperatives and Deontic Logic." *Analysis* 24 (2): 33–36. https://doi.org/10.1093/analys/24.2.33

Chisholm, Roderick M. and Sosa, Ernest. 1966. "On the Logic of 'Intrinsically Better.'" *American Philosophical Quarterly* 3 (3): 244–49.

Cohen, Philip R. and Levesque, Hector J. 1990. "Intention is Choice with Commitment." *Artificial Intelligence* 42 (2–3): 213–61. https://doi.org/10.1016/0004-3702(90)90055-5

D'Agostino, Marcello, Gabbay, Dov M., Hähnle, Reiner, and Posegga, Joachim. Eds. 1999. *Handbook of Tableau Methods*. Dordrecht: Kluwer Academic Publishers. https://doi.org/10.1007/978-94-017-1754-0

Downie, Robin S. 1984. "The Hypothetical Imperative." *Mind* (New Series) 93 (372): 481–90. https://doi.org/10.1093/mind/XCIII.372.481

Feldman, Fred. 1986. *Doing the Best We Can: An Essay in Informal Deontic Logic*. Dordrecht: D. Reidel Publishing Company. https://doi.org/10.1007/978-94-009-4570-8

Feldman, Fred. 2004. *Pleasure and the Good Life*. Oxford, New York: Oxford University Press. https://doi.org/10.1093/019926516X.001.0001

Firth, Roderick. 1952. "Ethical Absolutism and the Ideal Observer." *Philosophy and Phenomenological Research* 12 (3): 317–45. https://doi.org/10.2307/2103988

Fitting, Melvin, and Mendelsohn, Richard L. 1998. *First-Order Modal Logic*. Kluwer Academic Publishers. https://doi.org/10.1007/978-94-011-5292-1

Foot, Philippa. 1972. "Morality as a System of Hypothetical Imperatives." *The Philosophical Review* 81 (3): 305–16. https://doi.org/10.2307/2184328

Gabbay, Dov M. 1976. *Investigations in Modal and Tense Logics with Applications to Problems in Philosophy and Linguistics*. Dordrecht: Reidel. https://doi.org/10.1007/978-94-010-1453-3

Gabbay, Dov M., Horty, John, Parent, Xavier, van der Meyden, Ron, and van der Torre, Leendert. Eds. 2013. *Handbook of Deontic Logic and Normative Systems*. London: College Publications.

Garson, James W. 1984. “Quantification in Modal Logic.” In *Handbook of Philosophical Logic* 2 (2nd edition 3, 2001), edited by D. M. Gabbay and F. Guenthner, 249–307. Dordrecht: Springer. https://doi.org/10.1007/978-94-009-6259-0_5

Garson, James W. 2006. *Modal Logic for Philosophers*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511617737

Gensler, Harry J. 1985. “Ethical Consistency Principles.” *The Philosophical Quarterly* 35 (139): 156–70. https://doi.org/10.2307/2219341

Gensler, Harry J. 2002. *Introduction to Logic*. London and New York: Routledge.

Greenspan, Patricia S. 1975. “Conditional Oughts and Hypothetical Imperatives.” *The Journal of Philosophy* 72 (10): 259–76. https://doi.org/10.2307/2024734

Hansson, Sven Ove. 1990. “Defining ‘Good’ and ‘Bad’ in Terms of ‘Better.’” *Notre Dame Journal of Formal Logic* 31 (1): 136–49. https://doi.org/10.1305/ndjfl/1093635338

Harsanyi, John C. 1958. “Ethics in Terms of Hypothetical Imperatives.” *Mind* 67 (267): 305–16. https://doi.org/10.1093/mind/LXVII.267.305

Hill, Jr. Thomas E. 1973. “The Hypothetical Imperative.” *The Philosophical Review* 82 (4): 429–50. https://doi.org/10.2307/2183709

Hill, Jr. Thomas E. 1989. “Kant’s Theory of Practical Reason.” *The Monist* 72 (3): 363–83. https://doi.org/10.5840/monist198972320

Hilpinen, Risto. Ed. 1971. *Deontic Logic: Introductory and Systematic Readings*. Dordrecht: D. Reidel Publishing Company. https://doi.org/10.1007/978-94-010-3146-2

Hilpinen, Risto. Ed. 1981. *New Studies in Deontic Logic: Norms, Actions, and the Foundation of Ethics*. Dordrecht: D. Reidel Publishing Company. https://doi.org/10.1007/978-94-009-8484-4

Hilpinen, Risto, and McNamara, Paul. 2013. “Deontic Logic: A Historical Survey and Introduction.” In *Handbook of Deontic Logic and Normative Systems*, edited by D. M. Gabbay, J. Horty, X. Parent, R. van der Meyden and L. van der Torre, 3–136. London: College Publications.

Hintikka, Jaakko. 1961. “Modality and Quantification.” *Theoria* 27 (3): 117–28. https://doi.org/10.1111/j.1755-2567.1961.tb00020.x

Horty, John F. 2015. “Requirements, Oughts, Intentions.” *Philosophy and Phenomenological Research* 91 (1): 220–29. https://doi.org/10.1111/phpr.12204

Hughes, George E. and Cresswell, Max J. 1968. *An Introduction to Modal Logic*. London: Routledge. https://doi.org/10.4324/9780203028100

Jeffrey, Richard C. 1967. *Formal Logic: Its Scope and Limits*. New York: McGraw-Hill. https://doi.org/10.2307/2271990

Kant, Immanuel. 1991. *Grundlegung zur Metaphysik der Sitten*. In *The Moral Law: Kant’s Groundwork of the Metaphysics of Morals*. Translated and analysed by

H. J. Paton. London and New York: Routledge (Reprinted 1991; originally published 1785).

Kawall, Jason. 2013. "Ideal Observer Theories." In *The International Encyclopedia of Ethics*, edited by H. LaFollette, 2523–30. Blackwell Publishing. https://doi.org/10.1002/9781444367072.wbiee548

Knuuttila, Simo. 2004. *Emotions in Ancient and Medieval Philosophy*. Oxford: Oxford University Press. https://doi.org/10.1093/0199266387.001.0001

Korsgaard, Christine M. 2008. "The Normativity of Instrumental Reason." In *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*, Oxford/New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199552733.003.0002

Kracht, Marcus. 1999. *Tools and Techniques in Modal Logic*. Amsterdam: Elsevier.

Lenzen, Wolfgang. 1983. "On the Representation of Classificatory value Structures." *Theory and Decision* 15 (4): 349–69. https://doi.org/10.1007/BF00162113

Lewis, Clarence I., and Langford, Cooper H. 1932. *Symbolic Logic*. New York: The Century Company.

Lorini, Emiliano, and Herzig, Andreas. 2008. "A Logic of Intention and Attempt." *Synthese* 163 (1): 45–77. https://doi.org/10.1007/s11229-008-9309-7

Mally, Ernst. 1926. *Grundgesetze des Sollens: Elemente der Logik des Willens.* Leuschner and Lubensky.

Marra, Alessandra, and Klein, Dominik. 2015. "Logic and Ethics: An Integrated Model for Norms, Intentions and Actions." In *International Workshop on Logic, Rationality and Interaction*, edited by Wiebe van der Hoek, Wesley H. Holliday, and Wen-fang Wang, 268–81. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-48561-3_22

Marshall, John. 1982. "Hypothetical Imperatives." *American Philosophical Quarterly* 19 (1): 105–14.

Mele, Alfred R. Ed. 2004. *The Oxford Handbook of Rationality*. Oxford: Oxford University Press. https://doi.org/10.1093/0195145399.001.0001

Parks, Zane. 1976. "Investigations into Quantified Modal Logic I." *Studia Logica* 35 (2): 109–25.

Paton, Herbert J. 1948. *The Moral Law: Kant's Groundwork of the Metaphysics of Morals*. London and New York: Routledge (Reprinted 1991).

Priest, Graham. 2005. *Towards Non-Being*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198783596.001.0001

Priest, Graham. 2008. *An Introduction to Non-Classical Logic*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511801174

Prior, Arthur N. 1954. "The Paradoxes of Derived Obligation." *Mind* 63: 64–65.

Prior, Arthur N. 1958. "Escapism: The Logical Basis of Ethics." In *Essays in Moral Philosophy*, edited by Abraham I. Melden, 135–46. Seattle: University of Washington Press.

Rönnedal, Daniel. 2009. "Counterfactuals and Semantic Tableaux." *Logic and Logical Philosophy* 18 (1): 71–91. https://doi.org/10.12775/LLP.2009.006

Rönnedal, Daniel. 2012. "Bimodal Logic." *Polish Journal of Philosophy.* VI (2): 71–93. https://doi.org/10.5840/pjphil20126214

Rönnedal, Daniel. 2018. "Temporal Alethic Dyadic Deontic Logic and the Contrary-to-Duty Obligation Paradox." *Logic and Logical Philosophy* 27 (1): 3–52. https://doi.org/10.12775/LLP.2017.012

Rönnedal, Daniel. Forthcoming. "Contrary-to-Duty Obligations and the Contrary-to-Duty (Obligation) Paradox." *The Internet Encyclopedia of Philosophy.*

Ross, Alf. 1941. "Imperatives and Logic." *Theoria* 7: 53–71.

Ross, Alf. 1944. "Imperatives and Logic." *Philosophy of Science* 11 (1): 30–46. https://doi.org/10.2307/2268025

Schroeder, Mark. 2004. "The Scope of Instrumental Reason." *Philosophical Perspectives* 18 (1): 337–64. https://doi.org/10.1111/j.1520-8583.2004.00032.x

Schroeder, Mark. 2005. "The Hypothetical Imperative?" *Australasian Journal of Philosophy* 83 (3): 357–72. https://doi.org/10.1080/00048400500191958

Schroeder, Mark. 2009. "Means-End Coherence, Stringency, and Subjective Reasons." *Philosophical Studies* 143 (2): 223–48. https://doi.org/10.1007/s11098-008-9200-x

Schroeder, Mark. 2015. "Hypothetical Imperatives." In *Reason, Value, and Respect: Kantian Themes from the Philosophy of Thomas E. Hill, Jr.*, edited by Mark Timmons and Robert N. Johnson, Chapter 4. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199699575.003.0005

Semmling, Caroline, and Wansing, Heinrich. 2008. "From BDI and stit to bdi-stit Logic." *Logic and Logical Philosophy* 17 (1–2): 185–207. https://doi.org/10.12775/LLP.2008.011

Shaver, Robert. 2006. "Korsgaard on Hypothetical Imperatives." *Philosophical Studies* 129 (2): 335–47. https://doi.org/10.1007/s11098-004-1646-x

Smullyan, Raymond M. 1968. *First-Order Logic.* Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-642-86718-7

Sumner, Leonard W. 1996. *Welfare, Happiness, and Ethics.* Oxford: Clarendon Press. https://doi.org/10.1093/acprof:oso/9780198238782.001.0001

Sumner, Leonard W. 2000. "Something in Between." In *Well-Being and Morality: Essays in Honour of James Griffin*, edited by Roger Crisp and Brad Hooker, 1–19. Oxford: Clarendon Press.

van Roojen, Mark. 2013. “Internalism, Motivational.” In *International Encyclopedia of Ethics*, edited by H. LaFollette, 2693–706. Malden, MA: Wiley Blackwell.

von Wright, Georg H. 1968. *An Essay in Deontic Logic and the General Theory of Action*. Amsterdam: North-Holland.

Wallace, R. Jay. 2001. “Normativity, Commitment, and Instrumental Reason.” *Philosophers’ Imprint* 1 (3): 1–26.

Way, Jonathan. 2010. “Defending the Wide-Scope Approach to Instrumental Reason.” *Philosophical Studies* 147 (2): 213–33. https://doi.org/10.1007/s11098-008-9277-2

Organon F 27 (2) 2020: 263–279
https://doi.org/10.31577/orgf.2020.27204

ISSN 2585-7150 (online)
ISSN 1335-0668 (print)

RESEARCH ARTICLE

Horwich on the Value of Truth

Byeong D. Lee*

Received: 9 February 2019 / Accepted: 3 December 2019

Abstract: On the normativity objection to Horwich's minimalist theory of truth, his theory fails to capture the value of truth. In response to this objection, he argues that his minimalist theory of truth is compatible with the value of truth. On his view, the concept of truth is not constitutively normative, but the value of true beliefs can be explained instead by the belief-truth norm that we ought to want our beliefs to be true, and the value of true beliefs expressed in this norm is a moral value. I accept a deflationary theory of truth, according to which truth is too thin a concept to be constituted by any substantial norms. Thus I agree that the concept of truth is not constitutively normative. In this paper, however, I argue that the alleged value of true beliefs can be better explained in terms of epistemic normativity rather than moral normativity.

Keywords: Horwich; deflationism about truth; the value of truth; moral values; epistemic values.

1. Introductory remarks

According to Paul Horwich's minimalist theory of truth, the meaning of the truth predicate 'is true' is fixed by our underived acceptance of instances

* Sungkyunkwan University
✐ Department of Philosophy, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul, Republic of Korea
✉ bydlee@skku.edu

of the equivalence schema, ‘The proposition *that p* is true if and only if *p*.’ One important problem with this view is the so-called ‘normativity objection to minimalism.’ Many philosophers such as Michael Dummett (1959), Crispin Wright (1992), Hilary Putnam (1994a; 1994b), and Robert Brandom (1994) argue that our concept of truth is constitutively (or intrinsically) normative. On their views, we *ought* to speak and believe the truth, and so the concept of truth is to be understood in normative terms such as ‘what one *ought* to believe.’ But it seems that Horwich’s minimalism cannot capture this kind of *evaluative* character of truth. The reason is straightforward. Instances of ‘the proposition *that p* is true if and only if *p*’ merely tell us *when* beliefs possess the property of being true, and so these instances are completely silent on the question of whether its possession is *desirable or valuable*. Therefore, it seems that Horwich’s minimalism fails to capture the value of true beliefs (cf. Dummett 1959, 230–31; Brandom 1994, 17).

However, Horwich (2010; 2013) argues that his minimalist theory of truth is compatible with the value of truth. On his view, the concept of truth is not constitutively normative, but the value of true beliefs can be explained instead by the belief-truth norm that we ought to want our beliefs to be true, and the value of true beliefs expressed in this norm is a moral value. Many epistemologists take true beliefs as having a fundamental epistemic value rather than a moral value (see, e.g., Goldman 2001; Sosa 2001; Alston 2005; David 2014; Sylvan 2018). Thus, it is well worth examining whether true beliefs are indeed morally valuable.

I accept a deflationary theory of truth, according to which truth is too thin a concept to be constituted by any substantial norms (see Lee 2017). Thus, I agree with Horwich that the concept of truth is not constitutively normative. In this paper, however, I argue that the alleged value of true beliefs relevant to the normativity objection can be better explained in terms of epistemic normativity rather than moral normativity.

This paper proceeds as follows. In section 2, I introduce Horwich’s view that the value of true beliefs relevant to the normativity objection is moral. In section 3, I argue that the alleged value of true beliefs can be better explained in terms of epistemic normativity rather than moral normativity. Finally, in section 4, I address some possible objections that Horwich could raise against my alternative proposal.

2. Horwich's defense for the value of truth as a moral value

As pointed out before, on Horwich's minimalist theory of truth, the meaning of the truth predicate 'is true' is fixed by our underived acceptance of instances of the equivalence schema, 'The proposition *that p* is true if and only if *p*.' On this deflationary conception of truth, there is complete cognitive equivalence between the left-hand side of the biconditional and its right-hand side, and there is nothing else to say about truth other than what the truth predicate does; and the truth predicate serves only as a device of generalization, semantic ascent, and certain other logical or expressive functions, and so truth is not a substantial concept. In particular, on Horwich's view, the truth predicate is not an empirical predicate—such as 'red,' 'tree,' and 'magnetic'—which expresses a substantial, naturalistic property, but rather a logical predicate which expresses a non-substantial, logical property, which has no underlying nature (see Horwich 1998, 37; 2010, 15, 21, 31, and n. 23; and 2013, 25-26). To put the point another way, truth is too thin a concept to have an underlying nature, and so it is not the kind of thing that is constituted by some substantial norms.

How, then, does Horwich explain the alleged value of true beliefs? Horwich in his 2010 book titled *Truth-Meaning-Reality* argues that true beliefs are desirable, not because truth is constitutively normative, but because true beliefs are not only practically valuable but also non-instrumentally valuable. In his 2013 paper titled 'Belief-Truth Norms,' however, Horwich gives up the view that the value of true beliefs is partly due to instrumental desirability. He writes:

> First, it often happens that a person's true belief leads him to a decision that turns out *badly*, and he would have been better off with a false one (e.g. the man who dies from an operation that he correctly thought had a 99 per cent chance of success). And second, there are certain kinds of belief that appear to have no potential for practical import. [...] [Furthermore] there are various kinds of belief that we can be pretty sure will have no instrumental/pragmatic significance whatsoever. Think of certain views in metaphysics (e.g. that there's a plurality of concrete possible worlds), in esoteric areas of set theory (e.g. that every category

> has an appropriate Yoneda embedding), or in normative domains (e.g. that lying is wrong). Surely there can be no pragmatic explanation of why we should want our beliefs in these domains to be true. Assuming this to be so, we must conclude that the kind of desirability at issue in our general '*OUGHT*' norm isn't *instrumental* desirability. (Horwich 2013, 24)

What Horwich calls our general '*OUGHT*' norm is the following:

('*OUGHT*') We ought to want our beliefs to be true.

On his view, the value of true beliefs can be articulated by this norm, and the value of true beliefs expressed in this norm is moral; in other words, it is from a moral point of view that we ought to want our beliefs to be true. For example, he writes:

> Respect for truth is commonly recognized as *a virtue*. And this suggests that we regard the non-pragmatic value of truth as *moral*—that is, it's from a moral point of view that a person ought to want each of his beliefs to be true (including those whose truth could never promote the satisfaction of his desires). (Horwich 2013, 25)

And he continues to argue that the '*OUGHT*' norm is explanatorily fundamental, so that we cannot explain why it is true, although we can explain our commitment to this norm. Furthermore, he does not take this norm as an absolute norm. He writes:

> '*OUGHT*' does not aim to specify an *all-things-considered* obligation for us to want our beliefs to be true, but merely states one of the normative pressures on our belief-oriented desires—a presumably *epistemological* pressure. […] But the belief-truth '*OUGHT*'-norm is not thereby falsified; it remains valid *pro tem*, purporting to specify just one of the factors that bear on our overall appraisals. (Horwich 2013, 19)

In the remainder of this paper, however, I argue against Horwich's view that the value of true beliefs relevant to the normativity objection is moral.

3. The alleged value of true beliefs and our respect for truth

On Horwich's view, it is from a moral point of view that one ought to want one's beliefs to be true, and so we can explain the value of true beliefs as a moral value expressed in this belief-truth '*OUGHT*' norm. In this section, however, I argue that the alleged value of true beliefs relevant to the normativity objection can be better explained in terms of epistemic normativity rather than moral normativity.

First of all, all living animals need information about the world necessary for their survival; and they also have to do something in order to deal with some practical problem or other in their lives. We are no exception. Unlike mere animals, however, we are rational beings. On Kant's view (1996), it is our conception of ourselves that we are rational beings who can engage in theoretical and/or practical reasoning in order to determine what to believe and/or what to do. What then is our distinctive way of obtaining information about the world as rational believers? We acquire information about the world in a way that is bound by the norms of theoretical reason (or epistemic norms). In other words, unlike mere animals, we are by nature such rational beings whose beliefs are bound by the norms of theoretical reason. For example, our beliefs are bound by modus ponens. Thus, if you believe not only that if p then q but also that p, and if you care whether q, then you ought to believe that q. Of course, someone can believe in a way that violates some epistemic norm such as modus ponens. But unlike mere animals, such a person can be subject to rational criticism. In a similar way, we are rational beings whose actions are bound by the norms of practical reason (or practical norms) as well. For example, our actions are bound by the following means-end reasoning: if you ought to achieve end E, and if doing A is a means implied by your achievement of E, then you ought to do A. Due to this distinctive rational nature of ours, we engage in theoretical and/or practical reasoning in order to determine what to believe and/or what to do.[1]

Second, as pointed out before, on the deflationary view of truth, truth is too thin a concept to be constituted by any substantial norms, and so

[1] For a detailed discussion and defense of this view, see (Lee 2018).

there are no substantial norms of truth on the basis of which we can evaluate a belief as true or false. Besides, we cannot step outside our conceptual framework to judge whether a belief is true. At this point, it is important to recognize that, as Kant (1996) insists, it is our conceptual framework that provides the norms, criteria, or rules for defending (or criticizing) any claim. Therefore, it is inevitable to address any demand for justification on the basis of our conceptual framework. In other words, we have no other way but to rely on our conceptual framework to justify something. Hence, we have no other way but to evaluate whether a belief is true on the basis of our norms of epistemic justification. To put it another way, when we evaluate whether a proposition '*p*' is true, what we are really doing is to evaluate whether it is epistemically justified; and if it is epistemically justified, we can thereby assert (or believe) that it is true. In this regard, it is worth recalling the equivalence schema, according to which to say that *p* is equivalent to saying that '*p*' is true. In short, we have no other way but to evaluate whether a belief is true on the basis of our norms (or standards) of epistemic justification.

Third, epistemic justification is different from moral justification. How then can we distinguish between the two? As pointed out before, we are rational beings who can engage in theoretical and/or practical reasoning in order to determine what to believe and/or what to do. When we try to determine what to believe by engaging in theoretical reasoning, we take the *epistemic (or theoretical) point of view*, which is concerned with having true beliefs (and avoiding false beliefs). At this point, it is important to note that from the deflationist point of view, 'the claim that *p* is true' and 'it is a fact that *p*' are equivalent ways of expressing the same thing. Thus, we may say that when we try to determine what to believe from the epistemic point of view, we are concerned with having beliefs which reflect how the world really is. By contrast, when we try to determine what to do by engaging in practical reasoning, we take a *practical point of view*, which is concerned with bringing about what is desired or desirable. For this reason, epistemic justification and moral justification (as a species of practical justification) are fundamentally different kinds of justification. For the former is concerned with having beliefs which reflect how the world really is, whereas the latter is concerned with bringing about what is morally desirable.

On the basis of the above considerations, we may also distinguish between epistemic and moral values in the following way. As previously argued, we have no other way but to evaluate whether a belief is true on the basis of our epistemic norms. And if we can assert that a belief is justified in this way, we can also assert that it is true. Moreover, if one holds a belief in a way that violates an epistemic norm, then one can be subject to rational criticism. Along these lines, we may argue that to evaluate a belief as epistemically justified is to evaluate it positively from the epistemic point of view, whereas to evaluate a belief as epistemically unjustified is to evaluate it negatively from the epistemic point of view. And to evaluate a belief positively from the epistemic point of view is tantamount to taking it as having a positive epistemic value (or as being epistemically valuable). Similar remarks apply to moral values.

To begin with, we have no other way but to evaluate whether a certain thing is morally good on the basis of some relevant moral norms (or standards). In this context, it is worth considering Kant's famous claim that there is no moral goodness prior to and independent of the moral law. As he puts it, "the concept of good and evil must not be determined before the moral law […] but only […] after it and by means of it" (Kant 1996, 5:63). Therefore, when we evaluate whether a certain thing is morally good, what we are really doing is to evaluate whether it is morally justified on the basis of our moral norms; and if it is morally justified, we can thereby assert that it is morally good.[2] In addition, to evaluate a thing as morally good (or justified) is to evaluate it positively from the moral point of view. This is, in turn, tantamount to taking it as having a positive moral value (or as being morally valuable). On the basis of these considerations, we can argue that epistemic and moral values are different kinds of values.

There is one more thing to note. We have no good reason to think that the alleged value of true beliefs is a moral value. In this regard, three things are worth emphasizing. First, when we evaluate whether a proposition '*p*' is true, what we are really doing is to evaluate whether it is epistemically justified from the epistemic point of view, which is concerned with having beliefs which reflect how the world really is. By contrast, when we evaluate

[2] For a detailed discussion and defense of this view, see (Lee 2018).

whether a thing is morally good, what we are really doing is to evaluate whether it is morally justified from the moral point of view, which is concerned with bringing about what is morally desirable. Second, a moral property is supposed to be a substantial property. By contrast, the deflationary view of truth denies that truth is a substantial property; besides, there are too many trivial and unimportant true propositions which are not worth caring about. Third, the alleged moral value of a true belief does not play any significant role in Horwich's view except that it addresses the normativity objection to his minimalism.

If what I have argued so far are on the right track, there is no good reason to think that a true belief is morally valuable for its own sake; and the alleged value of true beliefs can be better explained in terms of epistemic normativity, namely that we ought to believe in accordance with our epistemic norms.

At this point, Horwich could retort that one's respect for truth is commonly recognized as a *moral* virtue. As I will argue in the remainder of this section, however, there is a better way to explain our alleged respect for truth. As has been emphasized, truth is a deflationary concept, so that we have no other way but to evaluate whether a belief is true on the basis of our epistemic norms. Therefore, we ought to determine what to believe in accordance with our epistemic norms. And our concern for true (or justified) beliefs is manifested by the fact that we determine what to believe in accordance with our epistemic norms and we do revise a belief of ours if it turns out to be unjustified. Let me illustrate this point.

Suppose that a truly evil person possesses a lot of knowledge about the world. Suppose also that he really cares about seeking knowledge, because knowledge is necessary for doing morally bad things in a clever way. Thus, we can say that he really wants his beliefs to be true. In this regard, it is worth noting that it is one thing to possess knowledge, and it is quite another to make use of knowledge for getting what one wants or desires. Suppose further that his knowledge includes the cure for an epidemic disease from which a lot of people are suffering. But he does nothing to save those people because he has no desire whatsoever to help others. After all, he is a truly evil person. In this case, he is morally blameworthy, especially because he cannot excuse himself on the grounds that he does not know the cure. And we could not regard him as displaying any moral virtue at all. Nonetheless, we could still regard him as displaying an epistemic virtue, because he really cares about seeking knowledge, and so he always holds beliefs in accordance with correct epistemic norms.

We can admit that we do care about having true beliefs. If what I have argued so far are correct, this is not because truth is morally valuable for its own sake. Our concern for true beliefs can be better understood in the following way. When we try to determine what to believe from the epistemic point of view, we are concerned with having true beliefs (that is, beliefs which reflect how the world really is). And we have no other way but to evaluate whether a belief is true on the basis of our epistemic norms. Thus, we (as rational believers) ought to believe in accordance with our epistemic norms. As a consequence, our concern for true (or justified) beliefs is manifested by the fact that we determine what to believe in accordance with our epistemic norms and we revise a belief of ours if it turns out to be unjustified.[3]

Here I do not mean to claim that my arguments in this section refute Horwich's view. Nevertheless, if what I have argued so far are on the right track, then our alleged respect for true beliefs can be better explained in terms of epistemic normativity rather than moral normativity.

4. Possible objections

In this final section, let me address some possible objections that Horwich could raise against my alternative proposal.

[3] Someone might motivate the claim that true beliefs are morally valuable in the following way. There are things which we morally ought to care about. And we need true beliefs to successfully deal with those things. Thus, we morally ought to care about having true beliefs. This line of argument is unavailable to Horwich, however. On this line of argument, true beliefs are valuable because they help us to bring about something else that is morally valuable, so that true beliefs are only instrumentally valuable. But there are many trivial true beliefs which have nothing to do with things which we morally ought to care about. More importantly, Horwich upholds the view that true beliefs are morally valuable for their own sake, rather than being instrumentally valuable.

In the previous section, I have argued that our concern for true beliefs is manifested by the fact that we determine what to believe in accordance with our norms of epistemic justification (henceforth, simply 'justification') and we revise a belief of ours if it turns out to be unjustified. Horwich has a different view on this matter, however. On his view, we should not tie our respect for true beliefs to our norms of justification. The first reason he gives is this. Imagine a community whose members deploy very different norms of justification from ours. The members of the community are convinced that their norms of justification promote true beliefs. And their concern for true beliefs is no less than ours. Unfortunately, however, their norms of justification are defective to the effect that most of their beliefs are not likely to be true. Even in such a case, it is from a moral point of view that they still ought to want their beliefs to be true. This line of thought suggests that their commitment to their norms of justification might not be best explained by their concern for true beliefs. And we could be in a similar situation as the members of this imagined community (see Horwich 2013, 28). As I will argue below, however, this kind of possibility does not pose a serious problem for the usual view that our concern for true beliefs is tied to our norms of justification.

Let us consider the aforementioned possibility that many of our norms of justification do not promote true beliefs, contrary to what we think. As argued in the previous section, we have no other way but to evaluate whether a belief is true on the basis of our norms of justification; if we can assert that it is justified in accordance with these norms, we can also assert that it is true; moreover, to evaluate a belief as justified is to evaluate it positively from the epistemic point of view, whereas to evaluate a belief as unjustified is to evaluate it negatively from the epistemic point of view. In addition to these, recall that when we try to determine what to believe from the epistemic point of view, we are concerned with having beliefs which reflect how the world really is. As a related point, recall also that from the deflationist point of view, 'the claim that *p* is true' and 'it is a fact that *p*' are equivalent ways of expressing the same thing. Accordingly, we should understand our epistemic aim of having true beliefs in a way that does not invoke a substantial concept of truth. One typical way of doing this is to understand our epistemic aim as that of determining, for any proposition '*p*,' whether *p*. To put it another way, our epistemic aim is to determine what to believe in such a way that our beliefs reflect how the world really is. And we have no other (rational) way but to evaluate whether *p* by evaluating whether '*p*' is justified. Along these lines, we can argue that epistemic justification is directly tied to our epistemic point of view, which is concerned with having true beliefs.

One more thing to note is that we can, at least in principle, evaluate any given norm (or standard) of justification in terms of whether it promotes true beliefs. Our assessments of justification are relative to the evidence available to us, and some contrary evidence might be available only in the future. Thus, a belief which is currently taken to be justified could lose its positive justificatory status later by virtue of some future evidence to the contrary. In addition, in a similar way that our beliefs can be subject to rational criticism, our norms of justification can be subject to rational criticism as well. As noted, a belief can lose its positive justificatory status if some relevant contrary evidence becomes available to us. In a similar vein, a norm of justification can lose its positive justificatory status if we come to have overwhelming reasons to think that it does not promote true beliefs. For this reason, if we are given some compelling reasons to think that a certain epistemic norm of ours does not promote true beliefs, we should give up the norm; and if we come up with a better epistemic norm for having true beliefs (that is, beliefs which reflect how the world really is), we can adopt it as our new norm of justification for the sake of promoting true beliefs.

If the above considerations are on the right track, Horwich's objection above does not pose a serious problem for the usual view that our concern for true beliefs is tied to our norms of justification.[4]

[4] Horwich (2010, Chapter 10, esp. 220-23) argues for what he calls the 'no theory' theory. On this theory, the correctness of our basic epistemic norms cannot be explained, roughly, for the following reason. We can explain less basic epistemic norms in terms of more basic ones. But we cannot repeat this process forever, and so, in the end, we are bound to reach the most basic epistemic norms, which are explanatorily fundamental; and we cannot explain the correctness of those truly-basic epistemic norms. It is beyond the scope of this paper to refute this theory. Thus, let me confine myself to briefly explaining why I do not accept it. As I have argued

But Horwich provides us with another reason against the usual view. He writes:

> Still, it may be thought that someone's concern for truth is revealed merely by there being *some* norms of justification that he respects. For the following explanation appears to hold no matter which particular constraint on belief is substituted for *C*:
>
> S believes that imposing *C* promotes truth.
> S wants his belief to be true.
> Therefore, S imposes *C*.
>
> But I would suggest that this explanation is defective, in that its desire-for-truth premise is redundant. For the first premise alone suffices to reach the conclusion. In other words: if S thinks that constraint *C* is truth-promoting, then we can already see why he imposes that constraint, without needing to assume, in addition, that he wants his beliefs to be true. [...] Thus it's a reasonable conjecture that our commitment to our familiar collection of doxastic constraints is neither explained by, nor a manifestation of, our respect for the value of truth. (Horwich 2013, 28)

elsewhere (Lee 2019a; 2019b), we can avoid the aforementioned regress problem by appealing to a coherence theory of justification. On the foundationalist theories of justification, the infinite regress of justification is impossible, and so we must admit that there are basic beliefs, which constitute a free-standing body of beliefs in the sense that they can justify other beliefs, but they are justified without recourse to other beliefs. Along the lines of a coherence theory of justification, however, we can argue that there are no such things as basic beliefs. Notice that even alleged basic beliefs such as perceptual beliefs are not exempt from being rationally criticized. For any belief, if it turns out that it does not help us to promote our epistemic aim, it can be rejected for the sake of our epistemic aim. A similar point applies to epistemic norms. The criteria for accepting an epistemic norm are not fundamentally different from the criteria for accepting a belief about the world. Hence, if it turns out that a certain epistemic norm of ours does not promote our epistemic aim, then we can reject or revise it. Along these lines, we may argue that for any epistemic norm, we can evaluate whether or not it is justified in a coherentist way. For a detailed discussion and defense of this view, see (Lee 2019a; 2019b).

On Horwich's view, if S believes that imposing constraint *C* promotes truth, then we can already see why he imposes *C*, without needing to assume the desire-for-truth premise. If S believes that imposing *C* promotes truth, then he will think and act in accordance with this belief. Admittedly, this role of the belief does not depend on the fact that S wants the belief to be true. However, a belief alone is not sufficient for generating an action. In this regard, it is important to recognize that imposing *C* is an action rather than a belief, and also that some relevant desire (or intention) is also required to generate an action. For example, if S wants a glass of water, and if he also believes that the clear liquid in the glass in front of him is water, then he will reach over to get the glass. If, however, S does not want a glass of water, we cannot expect that he will reach over to get the glass, even if he believes that the clear liquid in the glass is indeed water. A similar point applies to Horwich's claim above. Suppose that S does not want to promote true beliefs. In this case, we cannot expect that S will impose *C*, even if S believes that imposing *C* promotes true beliefs. Thus, consider the following alternative explanation:

> S ought to promote true beliefs. He can promote true beliefs only by imposing *C*. Therefore, S ought to impose *C*.

Suppose that S understands the validity of the above argument. Suppose also that he endorses its two premises. In this case, S will intend to promote true beliefs, and we can expect that he will impose *C*, because he believes that he can promote true beliefs only by imposing *C*. Thus, we can explain why S imposes *C*. Here notice that if S did not want to promote true beliefs in the first place, then he would not impose *C*, even if he believed that he can promote true beliefs only by imposing *C*. For this reason, S's desire (or intention) to promote true beliefs is not redundant for the explanation of why he imposes *C*. In addition, as has been emphasized, we have no other way but to evaluate whether a belief is true on the basis of our norms of justification. Consequently, we are justified in asserting (or believing) that *p* only when we have adequate reasons for '*p*'; and if we can assert that *p*, then we can also assert that '*p*' is true. Therefore, if we are justified in asserting that *p*, this is not because '*p*' happens to be true, but rather because we have adequate reasons for '*p*'. And we can rationally promote true beliefs only holding beliefs in accordance with our norms of justification. Moreover, we can express our concern for true beliefs by holding beliefs in this way. Hence, Horwich's second objection above also does not pose a threat to the usual view that our concern for true beliefs is tied to our norms of justification.

Finally, let me briefly consider whether what I have argued so far can be affected by Horwich's claim that the aforementioned belief-truth '*OUGHT*' norm is not an absolute norm. On his view, the '*OUGHT*' norm does not aim to specify an all-things-considered obligation, but instead it purports to specify just one contribution to the overall value of a belief. Consequently, the desirability of a true belief is not absolute, and so there can be circumstances in which a false belief is to be preferred on balance. Therefore, on Horwich's view, despite the fact that '*p*' is not true, it is possible that, all things considered, it is more valuable for S to believe that *p* than not to believe that *p*. To put it another way, a moral value for wanting one's belief to be true can be overridden by some pragmatic value for holding the belief.

To begin with, Horwich's claim that the '*OUGHT*' norm is not an absolute norm is problematic. Let us assume, for the sake of argument, that the value of true beliefs expressed in the '*OUGHT*' norm is moral, as Horwich insists. But moral values are presumably categorical values, which are valuable to each and every rational being, whereas pragmatic values are not categorical, because such values can vary from individual to individual. And it is widely accepted that moral values, which are categorical values, are not overridden by any pragmatic values, which are non-categorical values. Thus, Horwich owes us an explanation of why and how categorical values can be overridden by non-categorical values. The burden of proof in this case lies on his shoulders. In addition, if what I have argued in section 3 are on the right track, we have no good reason whatsoever to think that the value of true beliefs is moral, even if we grant that the '*OUGHT*' norm is not an absolute norm. For one thing, moral values are presumably substantial values, whereas the deflationary view of truth denies that truth is a substantial property. In particular, Horwich holds the view that truth is a sort of logical property. Thus, he owes us an explanation of why and how such a logical property can be morally valuable. The burden of proof in this case lies on his shoulders as well.

5. Concluding remarks

On the normativity objection to minimalism, Horwich's deflationary theory of truth fails to capture the value of truth. Horwich addresses this objection by arguing that truth is morally valuable for its own sake. If what I have argued in this paper are on the right track, however, we can better explain the alleged value of true beliefs (or our alleged respect for truth) in terms of epistemic normativity rather than moral normativity.

First, on the deflationary view of truth, truth is too thin a concept to be constituted by any substantial norms, and so there are no substantial norms of truth on the basis of which we can evaluate a belief as true or false. Thus, we have no other way but to evaluate whether a belief is true on the basis of the norms (or standards) of epistemic justification. To put it another way, when we evaluate whether a proposition '*p*' is true, what we are really doing is to determine whether it is epistemically justified; and if it is epistemically justified, we can thereby assert (or believe) that it is true. In this regard, it is worth recalling the equivalence schema, according to which to say that *p* is equivalent to saying that '*p*' is true.

Second, epistemic justification and moral justification (as a species of practical justification) are fundamentally different kinds of justification. For the former is concerned with determining *what to believe* for having beliefs which reflect how the world really is, whereas the latter is concerned with determining *what to do* for bringing about what is morally desirable.

Third, we care about having true beliefs. But this is not because truth is morally valuable for its own sake. Our concern for true beliefs can be better understood in the following way. When we try to determine what to believe from the epistemic point of view, we are concerned with having true beliefs. And we have no other way but to evaluate whether a belief is true on the basis of our epistemic norms. Thus, we (as rational believers) ought to believe in accordance with our epistemic norms. As a consequence, our concern for true (or justified) beliefs is manifested by the fact that we determine what to believe in accordance with our epistemic norms and we revise a belief of ours if it turns out to be unjustified.

Fourth, we have no good reason to think that the alleged value of true beliefs is moral. In this regard, it is noteworthy that the alleged moral value of true beliefs does not play any significant role in Horwich's view except that it addresses the normativity objection to his minimalism.

Along these lines, contrary to what Horwich claims, we can argue that the alleged value of true beliefs can be better explained in terms of epistemic normativity rather than moral normativity.

Acknowledgements

I'd like to thank an anonymous reviewer for *Organon F* for giving me helpful comments on an early version of this paper.

References

Alston, William. 2005. *Beyond Justification: Dimensions of Epistemic Justification*. Ithaca: Cornell University Press.

Brandom, Robert. 1994. *Making It Explicit*. Cambridge: Harvard University Press.

David, Marian. 2014. "Truth as the Primary Epistemic Goal: A Working Hypothesis." In *Contemporary Debates in Epistemology*, Second Edition, edited by Matthias Steup, John Turri, and Ernest Sosa, 363–77. Chichester: Wiley Blackwell.

Dummett, Michael. 1959. "Truth." *Proceedings of the Aristotelian Society* 59: 141–62. https://doi.org/10.1093/aristotelian/59.1.141

Goldman, Alvin. 2001. "The Unity of Epistemic Virtues." In *Virtue Epistemology: Essays on Epistemic Virtue and Responsibility*, edited by Abrol Fairweather, and Linda Zagzebski, 30–48. Oxford: Oxford University Press.

Horwich, Paul. 1998. *Truth*, Second Edition. Oxford: Clarendon Press.

Horwich, Paul. 2010. *Truth-Meaning-Reality*. Oxford: Clarendon Press.

Horwich, Paul. 2013. "Belief-Truth Norms." In *The Aim of Belief*, edited by Timothy Chan, 17–31. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199672134.003.0002

Kant, Immanuel. 1996. "Critique of Practical Reason." In *Practical Philosophy*, translated and edited by Mary J. Gregor, 139–271. Cambridge: Cambridge University Press.

Lee, Byeong D. 2017. "The Truth-Conduciveness Problem of Coherentism and a Sellarsian Explanatory Coherence Theory." *International Journal of Philosophical Studies* 25 (1): 63–79. https://doi.org/10.1080/09672559.2016.1236140

Lee, Byeong D. 2018. "The Moral Law as a Fact of Reason and Correctness Conditions for the Moral Law." *Dialogue: Canadian Philosophical Review* 57 (1): 47–66. https://doi.org/10.1017/S0012217317000622

Lee, Byeong D. 2019a. "A Kantian-Brandomian View of Concepts and the Problem of a Regress of Norms." *International Journal of Philosophical Studies* 27 (4): 528–43. https://doi.org/10.1080/09672559.2019.1632370

Lee, Byeong D. 2019b. "Gupta on Sellars's Theory of Perception." *Dialogue: Canadian Philosophical Review*. https://doi.org/10.1017/s0012217319000180

Putnam, Hilary. 1994a. "Does the Disquotational Theory of Truth Really Solve All Philosophical Problems?" In *Words and Life*, edited by James Conant, 264–78. Cambridge: Harvard University Press.

Putnam, Hilary. 1994b. "On Truth." In *Words and Life*, edited by James Conant, 315–29. Cambridge: Harvard University Press.

Sosa, Ernest. 2001. "The Place of Truth in Epistemology." In *Intellectual Virtues: Perspectives from Ethics and Epistemology*, edited by Michael DePaul and Linda Zagzebski, 155–80. Oxford: Oxford University Press.

Sylvan, Kurt. 2018. "Veritism Unswamped." *Mind* 127 (506): 381–435. https://doi.org/10.1093/mind/fzw070

Wright, Crispin. 1992. *Truth and Objectivity.* Cambridge: Harvard University.

Call for Papers

Value in Language

Guest editor: **Dan Zeman** (Slovak Academy of Sciences)

Many expressions in natural language are used to convey how we value parts of the world – things, events, actions, people. We use them to express our own opinions, but they also help us gain insight into what others think. Value and valuing is a crucial part of our lives: it guides us into action, it categorizes the world around us, it shapes our identity.

The special issue focuses on questions in the semantics of natural language expressions that are used to express value and valuing. Evaluative expressions (moral terms like "good," "bad" or "ought to," aesthetic adjectives like "beautiful," "ugly," "balanced," predicates of taste like "tasty," "disgusting," "boring," thick terms like "courageous" or "generous"), slurs like "boche" and expressives like "damn" are among the expressions that involve, in some way or another, valuing and value. Among the questions papers in the issue should address are the following:

- How do languages encode value (if at all)? – Should value be part of the semantics of a language or of pragmatics (or neither)?
- What are the best arguments for the main approaches to these expressions in the literature?
- How is disagreement involving the expressions in question to be accounted for?
- What is the connection between the semantics of these expressions and the social milieus in which they are used?
- How are the most prominent linguistic features of those expressions (e.g., the "hyper-projectivity" of slurs) to be treated?

The following authors have confirmed their contribution to the issue: **Bianca Cepollaro** (Vita-Salute San Raffaele University), **Stefano Predelli** (University of Nottingham), **Pekka Väyrynen** (University of Leeds).

Papers up to 7500 words (including references) tackling the questions above (but also others that might be of interest) should be send to submissions@organonf.com with the subject "ViL special issue" by JULY 15, 2020. 4-6 papers will be selected for publication after double-blind refereeing. The special issue is planned to come out in the spring of 2021.