compiled on 18/11/2023

Variations on a Montagovian Theme*

Wolfgang Schwarz

Published in *Synthese* 190 (2013): 3377–3395

Abstract. What are the objects of knowledge, belief, probability, apriority or analyticity? For at least some of these properties, it seems plausible that the objects are sentences, or sentence-like entities. However, results from mathematical logic indicate that sentential properties are subject to severe formal limitations. After surveying these results, I argue that they are more problematic than often assumed, that they can be avoided by taking the objects of the relevant property to be coarse-grained ("sets of worlds") propositions, and that all this has little to do with the choice between operators and predicates.

Note (2023): The published version of this paper contains a mistake: Clause (i) of "Theorem 5" is not an actual theorem. Thanks David Balcarras for pointing this out to me.

1 Introduction

Propositional attitudes such as knowledge and belief appear to have a subject and an object. The subject is the person who knows or believes; the object is that which is known or believed. But what kind of object is this? Two answers have been popular in the more systematic branches of epistemology and philosophy of mind. The first identifies objects of attitudes with something like sets of possible worlds. Sets of worlds capture the conditions under which the relevant attitude is true (or accurate or satisfied), and they come with a conveniently built-in boolean structure. On the other hand, cognitively different attitudes such as a belief that $2 + 2 = 4$ and a belief that there are infinitely many primes, can be true under the very same conditions, and it is tempting to represent this difference as a difference in content. This suggests modeling the objects of knowledge and belief as something more like a sentence: the sentence '$2 + 2 = 4$' is obviously different from the sentence 'there are infinitely many primes'.

Sentential accounts may also look attractive if our topic is not the modeling of doxastic or epistemic states, but the semantics of ordinary-language attitude reports. In English, for example, one may be able to truly assert 'Jones believes that 2+2=4' and yet deny 'Jones believes that there are

* Thanks to David Chalmers, Alan Hájek, Daniel Nolan, Dustin Tucker and two anonymous referees for helpful comments and discussion.

infinitely many primes'. The sentential functor 'Jones believes that —' therefore does not operate on the truth-conditions of the embedded sentence. For the present paper, I want to set this topic aside. I will be concerned with statements about propositional attitudes as they might occur in a systematic philosophical or scientific theory. The fact that attitude reports in English are hyperintensional does not refute theories that take the objects of attitudes to be sets of possible worlds. It merely follows that the connection between attitudes, as modeled by the theory, and ordinary-language attitude reports is not as tight as one might perhaps have thought.[1]

There is nothing special here about propositional attitudes. It is quite generally a good idea to express systematic theories in a regimented language that may deviate from ordinary usage e.g. by resolving ambiguities, minimising context-dependence and giving stipulative, precise definitions to theoretically important terms. This is common practice in science, and we should not shy away from it in philosophy.

The choice between sentences and sets of worlds also arises for other "modalities" besides propositional attitudes. Necessity and apriority, for example, have often been understood as either properties of sentences or properties of possible-worlds propositions. The choice is also well-known for probability, although here the sets of worlds are usually replaced by sets of more limited possibilities called 'outcomes'. *Via* probability, the issue arises all across science, from quantum mechanics and genetics to stochastic models in economics.

This paper is about a somewhat technical problem for the choice of sentences as the objects of a given modality. A classical form of the problem was presented in [Montague 1963], where Montague proved (in effect) that any reasonably powerful theory according to which some sentential predicate P satisfies the following three principles is inconsistent with basic arithmetic:

(1) $P(A) \supset A$;
(2) $P(P(A) \supset A)$;
(3) if $P(A)$, and A logically entails B, then $P(B)$.

Consider knowledge, understood as a property of sentences. Uncontroversially, whenever A is known, then A is true. At least for ideal subjects, it is also plausible that this fact itself may be known, and that knowledge is closed under logical consequence. Montague's result suggests that at least one of these assumptions must be false.

The topic of Montague's paper wasn't knowledge, but theories of provability or analyticity as expressed in the language of modal logic. Carnap and Quine had suggested that in modal logic, '$\Box A$' should be understood as saying that the embedded sentence A is provable or analytic. Montague argued that this interpretation would rule out the corresponding combination of (1)–(3) and thereby many familiar systems of modal logic. As [Skyrms 1978] demonstrates, this is not quite correct. One *can* interpret the box of modal systems like S4 and S5 as a property of sentences without running into Montague's problem. The reason is that the standard syntax of modal logic, which treats the box as a sentence operator rather than a predicate, is not powerful enough to run Montague's proof.

This fact may be reassuring for certain interpretations of modal logic, but I don't think it is of much help when we talk about theories of knowledge or probability – if only because such theories generally

1 In defense of this view, see, among many others, [Lewis 1986: 32–34], [Stalnaker 1991], and [Bach 1997].

can't be expressed in the standard language of modal logic. When we go beyond the syntactical constraints of standard modal logic, we have to face Montague's result. I will argue that this puts more serious pressure on sentential treatments of modalities than is commonly acknowledged.

I will begin with a more careful presentation of Montague's result. Montague's proof is closely related to earlier arguments due to Gödel, Tarski and Löb concerning the notion of provability in mathematical theories. Further results along similar lines have since been proved e.g. in [Thomason 1980b] and [Smullyan 1986], and I will add a few observations of my own. Afterwards, we will return to the question of what all this might tell us about knowledge, belief, apriority or probability.

2 Theme and Variations

Let me be a bit more precise than I have been so far. We will be concerned with theories of knowledge, belief, apriority or some other "modality". Until further notice, I assume that a theory of this kind can be identified with a logically closed set of sentences in a first-order language $\mathcal{L}$. As usual, I write $\vdash_T A$ to say that A is a member of the theory T. To bring out the inconsistency with basic arithmetic, I assume that $\mathcal{L}$ includes the standard vocabulary of arithmetic, so that it can express things like $2 + 2 = 4$.

A theory is *arithmetically sound* if it does not contain any arithmetical falsehoods (like $2 + 2 = 5$). The set L of logical truths, for example, is arithmetically sound, but it also contains very few arithmetical truths. A more informative theory is *Robinson Arithmetic*, also known as Q. Q can be axiomatised by five or six simple and uncontroversial statements about numbers, which can be found in any textbook on mathematical logic, see e.g. [Boolos et al. 2007: ch.16].[2] If a theory contains these axioms, it is called an *extension of Q*. A famous extension of Q is *Peano Arithmetic*, *PA*.

Most of our results will follow more or less directly from a certain fact about numerical relations, i.e. relations between natural numbers. To state this fact, we need some more terminology. A numerical relation R is called *recursive* if there is a mechanical algorithm to check, for any given numbers, whether they stand in this relation to one another or not. R is *recursively enumerable* if there is an algorithm for listing all and only the numbers that stand in the relation to one another. R is *weakly representable in* a theory T if there is a formula $A(x_1, \dots, x_n)$ such that, for all numbers $i_1, \dots, i_n$, $R(i_1, \dots, i_n)$ iff $\vdash_T A(i_1, \dots, i_n)$.[3] R is *strongly representable in T* if there is a formula $A(x_1, \dots, x_n)$ such that whenever $R(i_1, \dots, i_n)$, then $\vdash_T A(i_1, \dots, i_n)$, otherwise $\vdash_T \neg A(i_1, \dots, i_n)$. Here is the fact.

Lemma 1 (Representation Lemma)**.** *(i) Every recursive relation is strongly representable in any extension of Q. (ii) Every recursively enumerable relation is weakly representable in any arithmetically sound theory.*[4]

2 In standard formulations of Q, it is not made explicit that the quantifiers range only over numbers. The fact that every number has a successor, for example, is expressed as something like $\forall x \exists y(x + 1 = y)$. Since we are interested in theories that also talk about, say, persons, I assume that the restriction is officially made explicit. Thus $\forall x \exists y(x + 1 = y)$ turns into something like $\forall x(Nx \supset \exists y(Ny \wedge x + 1 = y))$. I will, however, generally leave the restriction implicit when I sketch proofs, since it tends to add a distracting layer of syntactic complexity.

3 By $A(i_1, \dots, i_n)$ I mean the formula that results from $A(x_1, \dots, x_n)$ by replacing each x_j with the standard numeral for the number i_j in $\mathcal{L}$.

4 See e.g. [Boolos et al. 2007: ch.16] for a proof of (i), and [Machover 1996: 258f.] for (ii).

The main use of Lemma 1 is to prove Lemma 2. To this end, we first associate, in some systematic manner, each expression A of $\mathcal{L}$ with a number $\ulcorner A \urcorner$ (using, for example, the code numbers of UTF-8). $\ulcorner A \urcorner$ is called the *Gödel number* of A. Lemma 1 then implies

Lemma 2 (Diagonal Lemma)**.** *If T is an extension of Q, then for any $\mathcal{L}$-formula $A(x)$ there is a sentence G such that $\vdash_T G \leftrightarrow A(\ulcorner G \urcorner)$.*

In other words, for any predicate $A(x)$ expressible in our language $\mathcal{L}$, there is a sentence G which "says of itself" that it satisfies the predicate, in the sense that, according to T, G is true iff $A(\ulcorner G \urcorner)$ is true. G is constructed as follows. Consider the function that maps any formula $\Phi(x)$ to its "diagonalisation" $\Phi(\ulcorner \Phi \urcorner)$. The corresponding function on Gödel numbers is recursive, and hence according to Lemma 1 strongly represented in any extension of Q by some formula $\text{DIAG}(x, y)$. Now given a formula $A(x)$, define G as the diagonalisation of $\exists y(\text{DIAG}(x, y) \wedge A(y))$. It is then easy to show that $\vdash_T G \leftrightarrow A(\ulcorner G \urcorner)$.

All sorts of interesting things follow immediately from these two lemmas. For a simple illustration, let T be a theory which, for some predicate $P(x)$, satisfies all instances of

(T) $\quad \vdash_T P(\ulcorner A \urcorner) \supset A$, and

(RN) $\quad$ If $\vdash_T A$ then $\vdash_T P(\ulcorner A \urcorner)$.

If T is an extension of Q, then by lemma 2 there is a sentence G such that $\vdash_T G \leftrightarrow \neg P(\ulcorner G \urcorner)$. Together with (T), this implies that $\vdash_T P(\ulcorner G \urcorner) \supset \neg P(\ulcorner G \urcorner)$; so $\vdash_T \neg P(\ulcorner G \urcorner)$. But then $\vdash_T G$, because $\vdash_T G \leftrightarrow \neg P(\ulcorner G \urcorner)$, and then $\vdash_T P(\ulcorner G \urcorner)$ by (RN). So T is inconsistent. Thus we've proven "Theorem 3" in [Montague 1963]:

Theorem 3. *If T is a consistent extension of Q, then it cannot satisfy all instances of (RN) and (T).*

To remove clutter, I will henceforth write $\Box A$ for $P(\ulcorner A \urcorner)$, and $\Diamond A$ for $\neg P(\ulcorner \neg A \urcorner)$. (RN) and (T) can then be written in a form that you may recognize from modal logic:

(T) $\quad \vdash_T \Box A \supset A$;

(RN) $\quad$ If $\vdash_T A$ then $\vdash_T \Box A$.

Here are some further principles that will play a role in what follows.

(K) $\quad \vdash_T \Box(A \supset B) \supset (\Box A \supset \Box B)$;

(D) $\quad \vdash_T \Box A \supset \Diamond A$;

(4) $\quad \vdash_T \Box A \supset \Box\Box A$;

(5) $\quad \vdash_T \Diamond A \supset \Box\Diamond A$;

(GL) $\quad \vdash_T \Box(\Box A \supset A) \supset \Box A$;

(RS) $\quad$ If $\nvdash_T A$ then $\vdash_T \neg\Box A$.

I write (□X) for the necessitated version of the corresponding principle (X), and (CX) for its converse, where applicable. For example, (□T) is $\vdash_T \Box(\Box A \supset A)$, and (CT) is $\vdash_T A \supset \Box A$. I say that a theory *satisfies* a schema like (T) or (RN) if all its instances are true if this theory is chosen as T.

The following bunch of theorems are proved much like theorem 3 above.[5]

Theorem 4 (Tarski's Theorem)**.** *If T is a consistent extension of Q, then it cannot satisfy (T) and (CT), nor (RN) and (RS).*

Theorem 5. *If T is a consistent extension of Q, then it cannot satisfy any of the combinations*

(i) ~~*(RN), (K), (4) and (5);*~~ *
(ii) (RN), (K), (4) and (D);
(iii) (RN), (K), (□4) and (D);
(iv) (RN), (K), (□T) and (C5).

Theorem 6 (Löb's Theorem, formalised)**.** *If T is an extension of Q and satisfies (RN), (K) and (4), then it satisfies (GL).*

As I mentioned in the introduction, the original investigations into the present territory concerned the notion of provability in formal mathematical theories. Since Peano Arithmetic, for example, has a recursive set of axioms, membership (equivalently, provability) in PA is recursively enumerable. By Lemma 1, it follows that there is an arithmetical predicate $P(x)$ that weakly represents, in PA, the property of being a (Gödel number of a) member of PA. That is, for all arithmetical sentences A,

$$(\mathrm{RN}^+) \quad \vdash_{\mathrm{PA}} A \text{ iff } \vdash_{\mathrm{PA}} P(\ulcorner A \urcorner).$$

The left-to-right direction of this is (RN). Since PA is consistent, Theorem 4 entails that no predicate *strongly* represents membership in PA. Hence there are sentences G such that $\nvdash_{\mathrm{PA}} G$ and $\nvdash_{\mathrm{PA}} \neg P(\ulcorner G \urcorner)$. By (RN^+), $\nvdash_{\mathrm{PA}} G$ entails $\nvdash_{\mathrm{PA}} P(\ulcorner G \urcorner)$, so PA can neither prove nor disprove $P(\ulcorner G \urcorner)$. This is Gödel's first incompleteness theorem. Investigating further into the predicate $P(x)$, one can show that PA satisfies (K) and (4). By theorem 6, this gives us (GL). Now suppose $\vdash_{\mathrm{PA}} \neg P(\ulcorner \bot \urcorner)$, equivalently: $\vdash_{\mathrm{PA}} \Box\bot \supset \bot$. By (RN^+), then $\vdash_{\mathrm{PA}} \Box(\Box\bot \supset \bot)$, so by (GL), $\vdash_{\mathrm{PA}} \Box\bot$, and then $\vdash_{\mathrm{PA}} \bot$ by (RN^+). This is Gödel's second incompleteness theorem: if PA can prove its own consistency, then it is inconsistent.

If you look back at our proof of Theorem 3, you can see that T does not have to be a purely arithmetical theory, nor does it have to be axiomatisable. More importantly, $P(x)$ (i.e., □) can be any predicate whatsoever, irrespective of whether it represents membership in T or whether it can be expressed in the basic language of arithmetic. The same is true for theorems 4–6. They apply to knowledge, belief or apriority in a philosophical theory T just as much as to provability in Peano Arithmetic. It might therefore be a bit misleading when Montague [1963] speaks of "syntactical"

∗ **Correction 2023:**I don't know why I thought this. It isn't true. There are consistent extensions of Q that satisfy (RN), (K), (4) and (5). Thanks to David Balcarras for pointing out the mistake.

5 See e.g. [Boolos et al. 2007] for proofs of theorems 4 and 6, and [Smullyan 1986] for results along the lines of Theorem

treatments of modality, which suggests that the relevant modality is defined syntactically, perhaps as provability in a certain deductive system.

In a philosophical context, (RN) may be problematic. If $\Box$ is interpreted as knowledge or belief, (RN) says that every sentence in our theory T is known or believed. Whether this is plausible depends on what else is in T. However, only a limited number of (RN) applications are needed to prove the above theorems, and one can generally recover the proofs by instead using necessitated versions of other principles, combined with (selected instances of)

($\Box$L) $\quad \vdash_T \Box A$, if A is a logical truth;

($\Box$Q) $\quad \vdash_T \Box A$, if A is an axiom of Q.

I write (QC) for the combination of ($\Box$L), ($\Box$Q) and (K). (QC) amounts to the claim that $\Box$ is closed under Robinson Arithmetical consequence. Here is an (RN)-free analog of Theorem 3 that brings us close to the result mentioned in the introduction.

Theorem 7. *If T is consistent, then it cannot satisfy (QC), ($\Box T$) and (T).*

Note that we no longer require T to be an extension of Q. Theorem 7 can be proved as follows. By the diagonal lemma, $\vdash_Q G \leftrightarrow \neg\Box G$. Hence $\vdash_L \mathsf{Q} \supset (G \leftrightarrow \neg\Box G)$, with L the set of logical truths and Q the conjunction of the axioms of Q. By ($\Box$L), then $\vdash_T \Box(\mathsf{Q} \supset (G \leftrightarrow \neg\Box G))$, and thus by (K) and ($\Box$Q), $\vdash_T \Box(G \leftrightarrow \neg\Box G)$. Now assume T satisfies ($\Box$T). Then $\vdash_T \Box(\Box G \supset G)$. Together with $\vdash_T \Box(G \leftrightarrow \neg\Box G)$, this entails $\vdash_T \Box\neg\Box G$ and $\vdash_T \Box G$ (using ($\Box$L) and (K)). If T also satisfies (T), we get $\vdash_T \neg\Box G$, which means that T is inconsistent.

If we suppose, in the last step, that T satisfies (4) instead of (T), we have $\vdash_T \Box\neg\Box G$ as well as $\vdash_T \Box\Box G$, which yields the main result of [Thomason 1980b] (although by a simpler route):

Theorem 8 (Thomason's Theorem)**.** *If T is consistent, then it cannot satisfy (QC), ($\Box T$), (4) and (D).*[6]

Along similar lines, it is not hard to verify the following.[7]

Theorem 9. *If T is consistent, then it cannot satisfy any of*

(i) (QC), ($\Box T$) and (C5);

5. 5.(iv) plays a central role in [Koons 1992].

6 Thomason actually does not include (D) in the statement of his theorem in [Thomason 1980b]. In his formulation, his theorem says that if T satisfies ($\Box$L), (K), ($\Box$T) and (4), then $\vdash_T \Box\mathsf{Q} \supset A$ for all formulas A. As indicated by the first sentence right after his proof sketch, however, he probably *meant* $\vdash_T \Box\mathsf{Q} \supset \Box A$. At any rate his proof sketch does not support the stronger conclusion. [Thomason 2011] contains an even more puzzling formulation of the theorem, using $\vdash_T \Box\mathsf{Q} \supset \Box A$ as the conclusion, but with a weaker version of (K) and with (C4) in place of (4). Since (C4) is a consequence of (QC)+($\Box$T), this is again strictly stronger then Theorem 8, and I don't know how it is supposed to be proved. Confusingly, Thomason refers to [Thomason 1980a] for further information about this alleged result, but as far as I can tell, the theorem is not even mentioned there.

7 Theorem 9.(i) follows immediately from the considerations in support of Theorem 7; for (ii)–(iv), it helps to first establish Theorem 10, which can be done by modifying the proof of Theorem 6 in the way we have just modified the proof of Theorem 3 to reach Theorem 7.

(ii) (QC), (□K), (4), (□4), (D) and (□D);
(iii) (QC), (□K), (4), (□4), (C4), (□C4) and (D);
(iv) (QC), (□QC), (C5) and (□C5).

Theorem 10. *If T satisfies (QC), (□K), (4) and (□4), then it satisfies (GL).*

Finally, let's prove the Montagovian result discussed in the introduction.

Theorem 11 (Montague's Theorem)**.** *If T is consistent with Q, then it cannot satisfy (□L), (K), (T) and (□T).*[8]

Here we have to make a somewhat different use of the diagonal lemma. Consider the function f that maps the Gödel number of any expression A to the Gödel number of $\mathsf{Q} \supset \neg A$. Since f is recursive it is represented in Q by some formula $\mathrm{F}(x, y)$. Applying the diagonal lemma to the predicate $\exists y(\mathrm{F}(x, y) \wedge P(y))$, we get a sentence G such that $\vdash_Q G \leftrightarrow \exists y(\mathrm{F}(\ulcorner G \urcorner, y) \wedge P(y))$. Moreover, $\vdash_Q \forall y(\mathrm{F}(\ulcorner G \urcorner, y) \leftrightarrow y = \ulcorner \mathsf{Q} \supset \neg G \urcorner)$. So $\vdash_Q G \leftrightarrow P(\ulcorner \mathsf{Q} \supset \neg G \urcorner)$. In other words, $\vdash_Q G \leftrightarrow \Box(\mathsf{Q} \supset \neg G)$. Hence $\vdash_L \mathsf{Q} \supset (G \leftrightarrow \Box(\mathsf{Q} \supset \neg G))$. By propositional logic, this entails $\vdash_L (\Box(\mathsf{Q} \supset \neg G) \supset (\mathsf{Q} \supset \neg G)) \supset (\mathsf{Q} \supset \neg G)$. Now suppose T satisfies (□T) as well as (□L) and (K). Then we get $\vdash_T \Box(\mathsf{Q} \supset \neg G)$. And since $\vdash_L \mathsf{Q} \supset (G \leftrightarrow \Box(\mathsf{Q} \supset \neg G))$, we then have $\vdash_T \mathsf{Q} \supset G$. If T also satisfies (T), this means that $\vdash_T \mathsf{Q} \supset G$ and $\vdash_T \mathsf{Q} \supset \neg G$, yielding $\vdash_T \neg \mathsf{Q}$: T is inconsistent with Q.

I went through an unusually large number of theorems, in part because different principles are plausible for different modalities. Montague's Theorem, for example, puts pressure on sentential treatments of knowledge, but not so much on treatments of thought or belief, for which condition (T) may be regarded as unrealistic. Another reason for giving so many results is to make clear that the problems are not generated by some particular principle such as (□T), which figures prominently in both Montague's and Thomason's Theorem. If we want to treat knowledge or belief as a predicate of sentences, we are not safe by giving up (□T). In light of Theorem 9, for example, it looks like we also have to give up the "positive introspection" principle (4) or its necessitation (□4).

Is there a general pattern behind our results? They all involve either a principle for "adding a box" – (RN), (CT), (4), (5) – or for "removing a box" – (T), (□T), (C4), (C5). Most, but not all, involve both. However, not every such combination is problematic. For example, the combinations (QC) + (T) + (4) and (RN) + (K) + (4) are known to be consistent, and indeed sound for provability in PA (the former with respect to complete arithmetic, the latter with respect to PA itself; see [Solovay 1976]).

None of the theorems above assume that the relevant modality is recursive, or recursively enumerable. The latter, however, is not implausible for knowledge, belief or provability. Suppose an agent has an arbitrarily large but finite number of "explicit" beliefs, and that we identify their total beliefs with everything that is a (first-order) logical consequence of the explicit beliefs. Then the set B of believed sentences is a recursively enumerable theory. By Lemma 1, membership in B is weakly represented in B itself by some purely arithmetical predicate $P(x)$, provided that the agent does not believe any arithmetical falsehoods. Much like in the case of provability in PA, it then follows that there are sentences G such that $\nvdash_B G$ and $\nvdash_B \neg\Box G$. Hence our theory of belief T cannot satisfy

8 This is closely related to "Theorem 1" in [Montague 1963]. Montague actually uses a slightly weaker version of (K), and assumes that *T extends* Q.

(5). Worse, suppose the agent believes the axioms of Peano Arithmetic. Then B is an axiomatisable extension of PA. One can then show that PA satisfies (RN), (K) and (4) with respect to the predicate $P(x)$.[9] If our theory of belief extends PA, it follows by theorem 5 that we can't even have (D), and by Theorem 6 that we must have (GL), and hence $\vdash_T \Box\neg\Box A \supset \Box A$. This suggests that there can be no sentence the agent doesn't believe of which they believe that they don't believe it!

Before we move on, it may be worth pointing out how the above results apply to probabilities. The simplest application is to read $\Box A$ as $P(A) = 1$. Standard probability theory guarantees that probability 1 is closed under logical consequence. Theorem 11, for example, thus suggests that either some false sentence must have probability 1 or there are sentences for which there is a positive probability that they have probability 1 despite being false: either $\vdash_T P(A) = 1 \wedge \neg A$ or $\vdash_T P(P(A) = 1 \wedge \neg A) > 0$. It follows that there can be no probability function that assigns 1 to all truths and 0 to all falsehoods. We can also prove some further facts, going beyond the general theorems above. For example, if the axioms of Q have probability 1, then by the diagonal lemma there is a sentence G such that $P(G \leftrightarrow P(G) < 1) = 1$. It follows that if $P(P(G) = 1) = 1$, then $P(G/P(G) = 1) = 0$, and if $P(P(G) = 1) < 1$, then $P(G/P(G) < 1) = 1$. Either way, the principle of *synchronous reflection*, $P(A/P(A) = x) = x$ must fail.

3 Dire Straits

In the previous section, we effectively used Gödel numerals as names for sentences: to say that a sentence A is known, we would say $P(\ulcorner A \urcorner)$, which applies a knowledge predicate $P(x)$ to the standard numeral of the Gödel number of A. An easy way to prevent the ensuing disasters is to use designated quotation instead. Suppose we introduce, for each sentence A, a new singular term 'A' that denotes A (rather than $\ulcorner A \urcorner$). The diagonal lemma still gives us a sentence G such that $\vdash_T G \leftrightarrow \neg P(\ulcorner G \urcorner)$, but we don't get $\vdash_T G \leftrightarrow \neg P(\text{'}G\text{'})$. So if we express our modal principles with designated quotation rather than Gödel numerals, we are safe.[10]

Unfortunately, this isn't much of a solution. First of all, there is a simple algorithm for translating between sentences and their Gödel numbers. Let *encode* be the function that turns a sentence into its Gödel number, and *decode* the inverse function that turns a Gödel number into the corresponding sentence. If our theory T can say a little bit about the syntax of sentences, it will inevitably have an expression $\text{DECODE}(x, y)$ that represents *decode*, in the sense that for any sentence A, $\vdash_T \text{DECODE}(\ulcorner A \urcorner, \text{'}A\text{'})$.[11] But then the diagonal lemma gives us e.g. a sentence G such that $\vdash_T G \leftrightarrow \exists x(\text{DECODE}(\ulcorner G \urcorner, x) \wedge \neg P(x))$. And then we simply re-run the proofs from the previous section with such slightly more complicated instances.

More generally, the arithmetical version of the diagonal lemma is just a particularly simple and

9 The reasoning is completely parallel to the reasoning that shows that PA satisfies (RN), (K) and (4) with respect to the predicate that represents provability in PA, see e.g. [Boolos 1993: ch.2].

10 See [Schweizer 1992: sec.3] for a detailed exposition of this strategy.

11 For example, with the Gödel numeration schema from [Boolos et al. 2007], $\text{DECODE}(x, y)$ can be defined from three predicates *Var*, *Pred* and *Char*, where *Var(q, n)* says that q is the n-th variable of $\mathcal{L}$, *Pred(q, n, m)* that q is the n-th m-ary predicate, and *Char(q, n, r)* that the n-th character of q is r.

well-known instance of a more general fact. Analogous lemmas can be proved in sufficiently strong versions of set theory, the lambda calculus, or theories of syntax. Identifying sentences with Gödel numbers is a convenient trick to stay in the well-charted territory of arithmetic; it is in no way essential to the formal results. We could just as well have taken a sufficiently strong theory of syntax as our basis, or a combination of minimal syntax and arithmetic, as in the previous paragraph.

Another initially plausible reaction to the results of section 2 starts from the observation that almost all of them assumed either (RN)+(K) or (□L)+(K), and thereby that the relevant modality is closed under logical consequence. Some modalities like knowledge and belief are arguably not closed under logical consequence.

Unfortunately, giving up closure does not help much either. One reason for this is pointed out in [Cross 2001b]. Let $P(x)$ stand for knowledge, and suppose it does not have the closure property (□L)+(K). Now extend $P(x)$ to (finite) lists of sentences so that $P(\ulcorner A_1, \ldots, A_n \urcorner)$ means that each of $A_1, \ldots, A_n$ is known. Then consider the *derivability* relation that holds between a list $A_1, \ldots, A_n$ and a sentence B iff $(A_1 \wedge \ldots \wedge A_n) \supset B$ is provable in first-order logic. The corresponding relation between Gödel numbers is recursively enumerable and thus weakly represented in any arithmetically sound theory T by some arithmetical predicate $\mathrm{DERIV}(x, y)$. Finally, define $P'(x)$ as $\exists y(P(y) \wedge \mathrm{DERIV}(y, x))$. $P'(x)$ expresses the property of being entailed by things that are known. We need not assume that this captures any intuitive sense of knowledge or "rational commitment". However, $P'(x)$ is guaranteed to be closed under logical consequence. So Theorem 11, for example, entails that it must violate either (T) or (□T) – blaming closure is no longer an option. But (T) and (□T) look just as plausible for $P'(x)$ as for $P(x)$.

Another reason why rejecting closure is of little help is that only a few specific instances of (□L) and (K) are used in the proofs of section 2. Instead of assuming that knowledge is in general closed under logical consequence, we can assume that our agent actually performed the relevant deductions, and satisfies whatever further conditions are required to know the conclusions. For example, in the proof of Theorem 7, we assumed that the agent knows the axioms of Q and inferred that they know the Q-theorem $G \leftrightarrow \neg\Box G$. Without closure, knowledge of the axioms does not entail knowledge of the theorem. But then imagine the agent just competently deduced $G \leftrightarrow \neg\Box G$ from the axioms of Q. In this case, surely, we might have knowledge of $G \leftrightarrow \neg\Box G$, and so we can continue with the proof of Theorem 7.

It is worth noting that nowhere in the previous section was it assumed that T consists of universal truths about the nature of the relevant modality. T may be any theory whatsoever. It may be a normative theory about knowledge, or a theory that describes the knowledge of some particular agent in some particular scenario. Whether or not knowledge is in general closed under logical consequence is therefore really beside the point. If knowledge always satisfies (T) and (□T), then in light of Montague's Theorem it can *never* be closed under logical consequence, nor can logical closure be a normative ideal.

This point also applies to the other principles. Consider the role of (□T) in Thomason's Theorem, whose intended application is to rational belief. Here (□T) amounts to the claim that the agent believes, for any sentence A, that they don't falsely believe A. There has been some discussion recently over whether this is a plausible constraint on rational belief. But this is the wrong question. If we

want to hold on to the other principles in Thomason's Theorem as general constraints on rational belief, we not only have to *allow* for agents who violate ($\Box$T), we have to *require* the violation. We have to say that there could not possibly be an agent whose rational beliefs satisfy ($\Box$T).

4 Retreat

Some of the results in section 2 can be reproduced using self-referential sentences of English. For example, consider the so-called "Knower sentence"

> This sentence is not known.

If the sentence is known, then it is false. So it is not known, for knowledge entails truth. This is just what it says. So we've found out that it is true. So now we know it. Contradiction. (Compare the proof of Theorem 3 on p.4.)

What went wrong? The first few steps seem unassailable: given that knowledge entails truth, the Knower sentence can't be known. Epistemic closure doesn't seem to play any objectionable role here. So we have to reject the last step, which uses an instance of ($\Box$T): while it is true that *if the Knower is known, then it is true (and hence not known)*, this fact itself cannot be known. But this is surely peculiar: we've just *proved* the allegedly unknowable fact!

At this point, the parallels to the Liar have certainly not escaped you. One of the results in section 2, theorem 4, actually *is* the formal version of the Liar paradox. It states that no modality can satisfy the truth-schema $\vdash_T \Box A \leftrightarrow A$. The proof appeals to a Liar-like sentence L which "says of itself" that it is not true, in the sense that $\vdash_T L \leftrightarrow \neg\Box L$.

A natural reaction to the Liar paradox is to declare the Liar sentence semantically defective. After all, if it is false, it would have to be true, and if it is true, it would have to be false, so it doesn't seem to have a unique, determinate truth-value. It is also well-known that various sensible procedures for assigning truth-values to sentences do not assign a stable truth-value to the Liar (see [Kripke 1975], [Gupta 1982], [Herzberger 1982]).

This points at a quite different way out: revise classical logic. Classical logic presupposes that every sentence is either true or false, and not both. If some sentence is neither true nor false, or both true and false, then certain principles of classical logic become problematic. The formal results of section 2 are still true, of course, but they assume that T is closed under classical consequence. If we only require closure in some non-classical logic, it is possible to find arithmetically competent theories T that satisfy e.g. the truth-schema $\vdash_T \Box A \leftrightarrow A$ without containing every sentence whatsoever.

How does this response to the Liar carry over to other modalities such as knowledge, belief, apriority or provability? The results in section 2 all employed the diagonal lemma to construct some kind of "self-referential" sentence. However, the arithmetical form of self-reference guaranteed by the diagonal lemma is semantically beyond reproach. Moreover, in contrast to the Liar, interpreting the diagonal sentences for knowledge or provability does not pose any immediate problems. If the relevant modality is recursively enumerable, the corresponding diagonal sentence is in fact a straightforward arithmetical equation. For provability in PA, this equation is even demonstrably *true* (and not false). Rejecting Gödel's Theorems on the grounds that the Gödel sentence is neither true nor false would be seriously misguided.

What about the Knower sentence? Since knowledge, like arithmetical provability, entails truth, this sentence can't be false. But there is no obvious *inconsistency* in assuming that it is true. On the other hand, suppose we want to model a conception of knowledge on which analytical truths are always known. More simply, suppose our topic is analyticity rather than knowledge. Since analyticity entails truth, the sentence that says of itself that it is not analytic cannot be false. So it is true, and true in virtue of its meaning. But then it is false, since it says that it is not true in virtue of its meaning. Contradiction. This suggests that the sentence that says of itself that it is not analytic, like the Liar and perhaps the Knower, is neither simply true nor simply false. So we have a reason to weaken classical logic and thereby escape the consequences of section 2. This kind of response to the Montagovian challenges for knowledge and belief has been defended e.g. in [Asher and Kamp 1989] and [Koons 1992].

I have no serious objections to these accounts. But we should not trick ourselves into thinking that there are special properties, perhaps including analyticity and knowledge, about which one must reason in a non-classical logic. If we stipulate that the knowledge predicate applies to arbitrary sentences of $\mathcal{L}$ and that it obeys ($\Box$L)+(K)+(T)+($\Box$T), we thereby preclude it from expressing any property. By a property I mean something that simply divides the relevant class of objects into those that have it and those that don't. If the Liar or the Knower doesn't have a determinate truth-value, this is because it involves a semantically defective predicate that does not correspond to a determinate property. Moving to a non-classical logic allows us to have a *word* 'knowledge' (or 'it is known that') that obeys the desired modal principles, but it does not vindicate the idea that there is a genuine property that satisfies these principles.

In section 1, I distinguished the ordinary use of words from their use in a systematic scientific, mathematical or philosophical theory. Various features of ordinary language, including vagueness, ambiguity and context-dependence, have been argued to stand in the way of assigning determinate truth-values to every sentence. We therefore shouldn't expect the logic of ordinary English to be classical. If an ambiguous sentence S is true on one reading and false on another, then S and its negation may count as both true, or as both untrue. However, if we want to develop a systematic theory on the topic of S, the best option is to disambiguate S into two sentences S_1 and S_2, rather than switching to a non-classical logic and saying that S is both true and false. Similarly, if it turns out that the English word 'knowledge' is semantically defective, the best choice for a systematic epistemology is to replace it with a better-behaved expression.

If we introduce an artificially regimented knowledge predicate into our theoretical language $\mathcal{L}$, the question arises whether the sentences that constitute the objects of knowledge are sentences of $\mathcal{L}$ or sentences of (say) ordinary English. In the latter case, knowledge directed at sentences involving the ordinary word 'knowledge' is not adequately formalised by a box within a box, because our theoretical knowledge predicate is different from the knowledge predicate in the "object language", i.e. the language that constitutes the objects of knowledge. We should then also reject the principle (T), for if A involves the object-language knowledge predicate, then $\Box A$ may be true in T, while A is not even a sentence of T. The results of section 2 are then no big surprise.

Some such separation between meta-language and object language is, to my mind, a satisfactory solution for semantic predicates like truth or analyticity as they might figure e.g. in a compositional

semantics for English. It is also relatively unproblematic for many uses of probability in science, as long as the relevant theory has no need to assign probabilities to statements which themselves involve probabilities. But in other cases, this solution looks quite radical. If knowledge, for example, is a relation to sentences, then it is hard to see why one couldn't in principle bear this relation to sentences involving a non-defective predicate for knowledge. Why couldn't we know the sentences of our own systematic epistemology?

Following Tarski, we could try to alleviate this problem by introducing a whole hierarchy of predicates. Sentences involving $\Box_0$ can never be objects of $\Box_0$, but they can still be objects of $\Box_1$. The problem with this is that the relevant phenomena usually don't sort themselves naturally into any such hierarchy. At what level of the belief hierarchy would we find, for example, my belief that some of my beliefs are false?

The best solution, in my view, for sentential accounts of propositional attitudes is to restrict the relevant modal principles. Again, this type of response is perhaps best known from the literature on truth. In Kripke's [1975: p.715] "closed off" construction, for example, the truth-schema $\vdash_T \Box A \leftrightarrow A$ is replaced by $\vdash_T (\Box A \vee \Box\neg A) \supset (\Box A \leftrightarrow A)$. The antecedent $\Box A \vee \Box\neg A$ excludes pathological sentences like the Liar. In the case of knowledge or belief, a simple but drastic restriction would limit the relevant principles to sentences that do not themselves contain a box. Various less drastic restrictions are also possible, see e.g. [Schweizer 1992: sec.4] and [des Rivieres and Levesque 1986]. Of course, this solution essentially means biting the bullet and accepting that ideal knowledge, for example, cannot obey ($\Box$L)+(K)+(T)+($\Box$T).

In this context, Löb's Theorem (Theorem 6) presents a challenge of its own. Recall that if we take $P(x)$ to represent the logical closure of an agent's beliefs, and the agent's beliefs form a recursively enumerable set that includes the axioms of Peano Arithmetic and no arithmetical falsehoods, then *it is provable in Peano Arithmetic* that $P(x)$ obeys (RN), (K) and (4) (with respect to PA). We are not given the choice to replace these principles by something weaker. Löb's Theorem then gives us (GL). But (GL) looks absurd if we read the box as belief: as mentioned above, it implies that there is no sentence which the agent doesn't believe of which they believe that they don't believe it. We should therefore reject logical closure as a general condition on belief. It still follows that an agent cannot believe that their beliefs are consistent (see Theorem 5.(ii)), and hence that there is no sentence for which it follows from the agent's beliefs that this sentence does not follow from the agent's beliefs. What if an agent *happens* to have logically closed beliefs? Then $P(\ulcorner\neg P(\ulcorner A\urcorner)\urcorner)$, where $P(x)$ is still defined as the logical closure of the agent's beliefs, does not adequately express the claim that the agent believes that they don't believe A, because the second P occurs in an intensional context. The agent can believe that they don't believe A without believing that A does not logically follow from something they don't believe. (If the agent *believes* that their beliefs are logically closed, then they really cannot believe that they disbelieve anything, as they cannot believe that they are consistent.)[12]

12 What if we read $P(x)$ as a predicate for what the agent *ought* to believe (in some sense), or as what beliefs would be warranted by their evidence (in some sense)? Plausibly, it is built into this notion that the relevant set of sentences is both consistent and logically closed. If it is recursively enumerable, it follows that there is no sentence of which the

5 Operators and predicates, sentences and propositions

In modal logic, combinations like (RN)+(T) or (RN)+(K)+(4)+(5) have been studied in great detail, and no-one doubts their arithmetical consistency. The main difference between these systems of modal logic and the systems discussed in section 2 is that modal logic treats the box as a sentence operator rather than a predicate of sentences. A simple and popular way to escape the Montagovian problems is therefore to formulate theories of knowledge, belief, necessity etc. with operators rather than predicates.

Philosophically, the distinction between operators and predicates looks a bit superficial. In standard models of modal logic, sentences are assigned truth-values relative to evaluation points ("worlds"); relative to a given point w, $\Box A$ is true iff A is true at all accessible points. The box in $\Box A$ therefore expresses a property of the set of points at which A is true, namely that it contains all accessible points. From a semantical perspective, the box *is* a predicate. This is even more obvious in provability logic, where $\Box A$ is read as saying that the sentence (assigned to) A is provable in a certain formal system. The merely orthographical difference between '$\Box A$' and '$P(A)$' or '$P(\text{`}A\text{'})$' is surely of no logical significance.

What, then, rescues modal systems like (RN)+(T) from the clash with basic arithmetic? Recall that to get our theorems going, we diagonalised a formula like $\exists y(\text{DIAG}(x, y) \wedge \neg P(y))$, which gave us a sentence G such that $\vdash_T G \leftrightarrow \neg P(\ulcorner G \urcorner)$. We can't do that in the standard language of modal logic, most obviously because we don't have quantifiers into the arguments of the box.

With this syntactical restriction in place, it is possible to give a sentential interpretation of modal operators without running into the results of section 2. This is nicely demonstrated in [Skyrms 1978] and [Cresswell 1985]. The basic idea of Skyrms and Cresswell resembles the "metalinguistic" interpretation of quantifiers used e.g. in [Machover 1996]. On this account, variables are treated like individual constants; $\forall x A(x)$ is true in a model M iff $A(x)$ is true in some model M' that differs from M at most in the interpretation of x. Since there are no assignment functions, the formula $A(x)$ is simply assigned a truth-value. To evaluate $\forall x A(x)$, we have to *re-interpret* the sentence $A(x)$ in different models. Skyrms and Cresswell show how to do the same with modal operators.[13]

So the lesson is that we can take the objects of (say) knowledge to be sentences of our theoretical language $\mathcal{L}$, and also endorse principles like (RN)+(T), as long as we never quantify over the objects of knowledge. But this is a bit like defending geocentric cosmology by refusing to look through telescopes. In ordinary talk, we commonly quantify over the objects of knowledge, and it is hard to accept that we must stop doing so in any systematic epistemology. We could never say that someone has come to know something, or that some person knows things that others don't.

Let's return to theories in which we can quantify over the objects of modalities. However, let's not assume that these objects are sentences of $\mathcal{L}$. What happens if the objects are sentences in a different language, or sets of possible worlds, or Russellian structured propositions?

Following the remarks at the beginning of section 3, I will assume, as I did with sentences of $\mathcal{L}$,

agent ought to believe that they ought not to believe it. This seems wrong. Presumably we should deny that the set of sentences the agent ought to believe is recursively enumerable.

13 Skyrms writes '$*Q(A)$' instead of '$\Box A$', and suggests that $Q(A)$ may be read as the quotation of A, and $*$ as a sentential predicate. Paradox returns if one adds quantifiers into the position of $*$; see [Schweizer 1987].

that $\Box$ is actually a predicate of numbers which somehow encode the actual objects of the modality. Now suppose, for a simple illustration, that we read $\Box A$ as attributing a certain property not to the syntactical entity A, but to its *truth-value*. We can encode truth-values by the standard Gödel numbers 0 (false) and 1 (true). For concreteness, let the box express the property of being equal to 1. $\Box A$ can then be expressed in the language of arithmetic as $\langle A \rangle = 1$, where $\langle A \rangle$ is the standard numeral for the truth-value of A. It is easy to verify that this renders all instances of ($\Box$L), (K), (T) and ($\Box$T) true. For example, since all logical truths are true, every instance of ($\Box$L) reduces to the single sentence $1 = 1$, which is true. For another example, every instance of (T) has either the form $1 = 1 \supset A$, where A is true, or $0 = 1 \supset A$, where A is false. Again, these are all true. So, on this reading of the box, first-order arithmetic satisfies the Montagovian combination ($\Box$L)+(K)+(T)+($\Box$T), and it does so without any artificial syntactical limitations.

Why does this reading of the box escape Montague's result? Couldn't we use the "decode" trick from section 3 again? Let's try. Let *val* be the function that maps (the Gödel number of) a sentence to (the Gödel number of) its truth-value. Let T be an arbitrary theory, and suppose *val* is (strongly) represented in T by some formula $\text{VAL}(x, y)$. Then we can re-run all the proofs of section 2, using e.g. the diagonal sentence G such that $\vdash_T G \leftrightarrow \exists x(\text{VAL}(\ulcorner G \urcorner, x) \wedge \neg P(x))$ where we previously used the sentence G with $\vdash_T G \leftrightarrow \neg P(\ulcorner G \urcorner)$. Here, *val* plays a parallel role to *decode* in section 3. However, Lemma 1 guarantees that the *decode* function is represented in any reasonably strong theory T, because *decode* is recursive. This is not true for *val*. In fact, the present considerations show that *val* is not recursive, and, more strongly, not representable in any arithmetically sound theory. (A more familiar argument to this conclusion goes via Tarski's Theorem.)

The situation looks similar if we replace truth-values with sets of possible worlds. Since arithmetical facts are not contingent, all arithmetical sentences express either the empty set or the set of all worlds. Encode these two sets as the numbers 0 and 1. If $\mathcal{L}$ can also express contingent facts, assign arbitrary numbers to the other sets of worlds expressible in $\mathcal{L}$. Again it is easy to show that the function that maps (the Gödel number of) every sentence to (the Gödel number of) the corresponding set of worlds is not representable in any arithmetically sound theory.

On the other hand, suppose we take the objects of modalities to be sentences of some language $\mathcal{L}'$, which may or may not be identical to $\mathcal{L}$. So $\Box A$ attributes a certain property (being known, for example) to a certain sentence of $\mathcal{L}'$. Now presumably there is some kind of systematic relationship between the sentence A here and the corresponding $\mathcal{L}'$-sentence that constitutes the object of the modality. More precisely, there will *some* encoding of $\mathcal{L}'$-sentences such that the relation between Gödel numbers of $\mathcal{L}$-sentences and the Gödel numbers of the corresponding $\mathcal{L}'$-sentences is recursive. But then it follows by Lemma 1 that this relation is (strongly) represented in every theory that comprises a minimum of arithmetic. And thus we can re-run the proofs in section 2.

Similarly for other candidate objects of modalities. In particular, suppose we introduce "impossible worlds" to the possible-worlds framework, so that we can assign different sets of worlds to different logical truths. Most simply, this could be done by defining worlds as arbitrary sets of $\mathcal{L}$-sentences. Then we could assign to any sentence S the set of "worlds" that contain S. But now encode every such set by the Gödel number of the corresponding sentence S. The relation between Gödel numbers of sentences and Gödel numbers of the sets of worlds expressed by those sentences is then represented in

basic arithmetic by the identity predicate, which brings back all the Montagovian results. More generally, if enriching the space of worlds is supposed to help with the "problem of logical omniscience", there will presumably be some encoding relative to which the relation between $\mathcal{L}$-sentences and the corresponding sets of worlds is recursive.[14] It is not enough to allow that different $\mathcal{L}$-sentences can express the same set of worlds, as long as there is an algorithm for deciding which sentences express the same set of worlds. To avoid the Montagovian results, there would have to be no encoding relative to which the relation between sentences and sets of worlds is recursive, or recursively enumerable, or even just representable in complete first-order arithmetic.

The fact that all logical and mathematical truths express the same possible-worlds proposition is often considered to be the main weakness of possible-worlds account. Here it is precisely this feature that saves the day.

6 Revenge?

It is sometimes pointed out, following [Asher and Kamp 1989], that possible-worlds treatments do not ultimately avoid the Montagovian results, because the problems return as soon as we add to our language a predicate for the *expression* relation between sentences and propositions. The fact that this relation is non-arithmetical in the possible-worlds framework only means that it is not automatically represented in any theory of sufficient arithmetical power. But we may still be tempted to add a new, non-arithmetical predicate to our language that represents this relation. And that is all we need to re-run the arguments of section 2.

However, the temptation to add an expression predicate should be resisted. As mentioned above, the expression relation is not representable in any arithmetically sound theory. If we add the predicate to a theory of sufficient arithmetical strength, we only render the theory inconsistent. We shouldn't do that.

The problem can be illustrated with a self-referential sentence of English. Consider the following sentence – call it E:

> This sentence (actually) expresses the empty set.

Suppose E expresses the empty set and hence is true. Then the actual world is a member of the proposition expressed by E. But the actual world isn't a member of the empty set. So E is false. This means that E is true at some world w. But then it is true at w that E expresses the empty set at the actual world, which entails that E actually expresses the empty set. Contradiction.

So it is true that a possible-worlds account of, say, knowledge is threatened by paradox if one adds an expression predicate. But this is because *every theory whatsoever* is threatened by paradox if one adds such a predicate. To be sure, it would be convenient if we could have a well-behaved, full-fledged expression predicate in a possible-worlds account. Since we can't, we have to make do with a stratified predicate that is not applicable to sentences containing itself, or with a predicate that does

14 Some impossible-worlds accounts only aim at modeling inconsistent belief, without addressing the problem of logical omniscience (e.g. [Restall 1997], [Berto 2010]). Depending on the details, these accounts may not be subject to the Montagovian limitations.

not represent the expression relation for pathological sentences like E. Either choice is, I think, good enough for a satisfactory theory of knowledge or belief.

[Koons 1992] argues that the Montagovian paradoxes, at least for rational belief, do not depend on a sentential treatment, because one can find empirical substitutes for the sentence G that figures in the proofs of section 2. It is easy to find instances of $G \leftrightarrow \neg\Box G$: any true sentence which the agent does not rationally believe will do. But $\vdash_T G \leftrightarrow \neg\Box G$ by itself is harmless. We typically also need the necessitated $\vdash_T \Box(G \leftrightarrow \neg\Box G)$. Here I do not find Koons's examples very convincing. They mostly involve things like an action that is "irrational if and only if it is not irrational" (p.17), which strikes me as obviously impossible.

I do not want to rule out the possibility of epistemic dilemmas: situations in which some epistemic norm entails that an agent ought to believe a proposition A, while another (or even the same) norm entails that they must not believe A. If there are such dilemmas, this would raise interesting issues for the logic of rational belief. Among other things, it would show that principle (D) is problematic if the box is read as "ought to believe". It might also show that an agent's epistemic obligations are not closed under logical consequence – otherwise the fact that they ought to believe A and also ought to not believe A would seem to entail that they ought to make true every proposition whatsoever. Nevertheless, these issues are quite different from those raised by the Montagovian results.

[Prior 1961] also discusses paradoxes for propositional attitudes that do not rely on a sentential treatment. Suppose Mr.X, who believes himself to be in room 6, thinks that nothing presently thought in room 7 is true. Suppose further that Mr.X is in room 7, and nothing else is thought in room 7 at that time. Then it looks like what Mr.X thinks is true iff it is not true. Prior concludes that the scenario of Mr.X is impossible.

Let's investigate this case from the perspective of a possible-worlds account. Let R be the relation that holds between a possible world w and a set of worlds x iff x is the content of a thought in room 7. In the world of Mr.X, call it w^*, the only thought in room 7 is supposed to have the content $p = \{w : \forall x(wRx \supset w \notin x)\}$. It follows that $w^* \in p$ iff $\forall x(w^*Rx \supset w^* \notin x)$ iff $w^* \notin p$. Contradiction. Set theoretically, the status of p is impeccable, assuming that R is a genuine relation between worlds and sets of worlds. In general, for any relation R between (set-theoretic) individuals and sets of individuals, there is a set of individuals $p = \{y : \forall x(yRx \supset y \notin x)\}$. However, this definition ensures that no individual stands in R to p and to nothing else. The paradox arises when we think of p as some fixed set, and consider the hypothesis that some individual stands in R exclusively to this set. Consider a simpler analogy. Let q be a set of individuals, and define p as $\{x : x \notin q\}$. For any choice of q, there is a corresponding set p. But although we are completely free to choose any set of individuals as q, and although p will always be a set of individuals, we can't stipulate that q should be the set of individuals p. Similarly in the case of Mr.X: since p is defined as $\{w : \forall x(wRx \supset w \notin x)\}$, there can be no world at which Mr.X is thought-related only to p. On the possible-worlds account, the scenario is indeed impossible. It *seems* possible, because we are tempted to think of p as some fixed set of worlds (perhaps as whatever proposition Mr.X would express by uttering "nothing presently though in room 7 is true") and then wonder why Mr.X couldn't be thought-related to this proposition.

There is a lot more to be said about these puzzles. But we shouldn't take it for granted that all paradoxes involving truth, belief or knowledge are of a single kind. Prior's puzzles, for example, do

not require any noteworthy modal principles at all. Hence most of the strategies I have explored in sections 4 and 5 are of no use here. In the other direction, a promising solution to Prior's puzzles – explaining how a systematic conception of the relevant modality renders the paradoxical situation impossible – is of no use for the Montagovian problems, since these arise independently of any empirical assumptions. Moving to sets of worlds helps to avoid the Montagovian limitations. It does not magically solve every other puzzle and paradox along the way.

I am not suggesting that we should in general take sets of worlds (or truth-values) to be the objects of modalities. Provability and syntactical well-formedness are clearly properties of sentences. There are also properties in the vicinity of knowledge and belief that uncontroversially apply to sentences. A classic example is Carnap's explication of belief in terms of dispositions to assent. All these properties are subject to the results of section 2. Sometimes this is not a problem. Syntactical well-formedness, for instance, clearly obeys none of (T), (C5) or (D), and clearly does obey (GL). In other cases, I think the formal limitations are a significant cost. For example, it is hard to accept that a systematic theory of apriority would have to give up one of ($\Box$L), (K), (T) and ($\Box$T). In fact, it is natural to think that the logic of apriority should be S5, at least on some way of cashing out this notion. But then apriority cannot be a property of sentences.[15] Similarly, one might want the logic of ideal knowledge to be S4 (or at any rate an extension of T), and one might want to have a notion of probability that makes synchronic reflection at least formally possible, and that allows for "maximally accurate" probability functions which assign 1 to every truth and 0 to every falsehood. In these cases, it may be preferable to identify the objects of the relevant modality with something like sets of possible worlds.

Some authors regard theories of knowledge, belief or probability as if they were concerned with a determinate, special property (*knowledge*! *belief*!) with which we are all intimately acquainted. But there are really a lot of properties out there, many of which deserve to be studied. Consider belief. Almost everyone agrees that beliefs can be evaluated for truth and falsity, and that they can be evaluated not just at the actual world, but also under merely hypothetical conditions. This means that for every belief there is an associated set of possible circumstances ("worlds") at which the belief is true. It should therefore be uncontroversial that wherever there is belief, there is a relation to sets of possible worlds. This relation is the topic of possible-worlds accounts. When people object to such accounts, they rarely give any reason to think that this relation does not exist or that it is not worth studying. Their objections rather indicate that they are thinking of a different relation, perhaps a relation between agents and certain structures in their "belief box", or a relation that holds between an agent and a sentence S iff one can truly say in ordinary English that the agent "believes S". The only competition here lies in the theoretical virtues of theories dealing with these different relations.

References

Nicholas Asher and Hans Kamp [1989]: "Self-Reference, Attitudes and Paradox". In G. Chierchia, B.H. Partee and R. Turner (Eds.) *Properties, Types and Meaning. Vol.1: Foundational Issues,*

15 That apriority satisfies S5 is suggested by the "two-dimensional semantics" of [Chalmers 2004], where apriority is modeled as truth at all worlds considered as actual. Yet according to [Chalmers 2004: 181f.], apriority is a property of sentences or "thoughts"; this suggests that the logic of apriority must actually be weaker than T (by Theorem 3).

Dordrecht: Kluwer, 85–158

Kent Bach [1997]: "Do Belief Reports Report Beliefs?" *Pacific Philosophical Quarterly*, 78: 215–241

Francesco Berto [2010]: "Impossible Worlds and Propositions: Against the Parity Thesis". *The Philosophical Quarterly*, 60: 471–486

George Boolos [1993]: *The Logic of Provability*. Cambridge: Cambridge University Press

George Boolos, John Burgess and Richard Jeffrey [2007]: *Computability and Logic*. New York: Cambridge University Press, 5th edition

David Chalmers [2004]: "Epistemic two-dimensional semantics". *Philosophical Studies*, 118(1-2): 153–226

Mac J. Cresswell [1985]: "We are all children of God". In B. K. Matilal and J. L. Shaw (Eds.) *Analytical Philosophy in Comparative Perspective,* Dordrecht: Reidel, 39–60

Charles B. Cross [2001a]: "The Paradox of the Knower Without Epistemic Closure". *Mind*, 110: 319–333

— [2001b]: "A Theorem Concerning Syntactical Treatments of Nonidealized Belief". *Synthese*, 129(3)

— [2004]: "More on the Paradox of the Knower Without Epistemic Closure". *Mind*, 113: 109–114

Jim des Rivieres and Hector J. Levesque [1986]: "The Consistency of Syntactical Treatments of Knowledge". In *Proc. of the 1986 Conference on Theoretical Aspects of Reasoning about Knowledge,* Monterey, CA, 115–130

Paul Égré [2005]: "The Knower Paradox in the Light of Provability Interpretations of Modal Logic". *Journal of Logic, Language and Information*, 14: 13–48

Anil Gupta [1982]: "Truth and Paradox". *Journal of Philosophical Logic*, 11: 1–60

Hans Herzberger [1982]: "Notes on Naive Semantics". *Journal of Philosophical Logic*, 11: 61–102

James M. Joyce [2007]: "Epistemic Deference: The Case of Chance". *Proceedings of the Aristotelian Society*, 107/2: 187–206

David Kaplan and Richard Montague [1960]: "A paradox regained". *Notre Dame Journal of Formal Logic*, 1: 79–90

Robert C. Koons [1992]: *Paradoxes of belief and strategic rationality*. Cambridge: Cambridge University Press

Saul Kripke [1975]: "Outline of a Theory of Truth". *Journal of Philosophy*, 72: 690–716

David Lewis [1986]: *On the Plurality of Worlds*. Malden (Mass.): Blackwell

Moshe Machover [1996]: *Set Theory, Logic and their Limitations*. Cambridge: Cambridge University Press

Richard Montague [1960]: "Logical necessity, physical necessity, ethics and quantifiers". *Inquiry*, 4: 259–269. Reprinted in [Montague 1974]

— [1963]: "Syntactical Treatments of Modality, with Corollaries on Reflection Principles and Finite Axiomatizability". *Acta Philosophica Fennica*, XVI: 153–167. Reprinted in [Montague 1974: 286–302].

— [1974]: *Formal Philosophy*. New Haven: Yale University Press

Richard Otte [1982]: "Modality as a metalinguistic predicate". *Philosophical Studies*, 41: 153–159

Arthur Prior [1961]: "On a Family of Paradoxes". *Notre Dame Journal of Formal Logic*, 2: 16–32

Willard van Orman Quine [1953]: "On a So-Called Paradox". *Mind*, 62: 65–67

Greg Restall [1997]: "Ways Things Can't Be". *Notre Dame Journal of Formal Logic*, 38: 583–597

Paul Schweizer [1987]: "Necessity viewed as a semantical predicate". *Philosophical Studies*, 52: 33–47

— [1992]: "A Syntactial Approach to Modality". *Journal of Philosophical Logic*, 21: 1–31

R. Shaw [1958]: "The Paradox of the Unexpected Examination". *Mind*, 67: 382–384

Brian Skyrms [1978]: "An Immaculate Conception of Modality, or How to Confuse Use and Mention". *Journal of Philosophy*, 75: 368–387

Raymond M. Smullyan [1986]: "Logicians who reason about themselves". In *Proceedings of the 1986 TARK Conference, Ed. J.Y.Halpern,* Morgan Kaufmann, 341–352

Robert Solovay [1976]: "Provability Interpretations of Modal Logic". *Israel Journal of Mathematics*, 25: 287–304

Robert Stalnaker [1991]: "The Problem of Logical Omniscience I". *Synthese*, 89

Richmond H. Thomason [1980a]: "A Model Theory for Propositional Attitudes". *Linguistics and Philosophy*, 4: 47–70

— [1980b]: "A Note on Syntactical Treatments of Modality". *Synthese*, 44

— [2011]: "Some Limitations to the Psychological Orientation in Semantic Theory". *Journal of Philosophical Logic*, 40: 1–14

Gabriel Uzquiano [2004]: "The Paradox of the Knower Without Epistemic Closure?" *Mind*, 113: 95–107