Modality and Expressibility*

Matthew Mandelkern†

December 17, 2018

Penultimate draft; to appear in the *Review of Symbolic Logic*

Abstract

When embedding data are used to argue against semantic theory A and in favor of semantic theory B, it is important to ask whether A could make sense of those data. It is possible to ask that question on a case-by-case basis. But suppose we could show that A can make sense of *all* the embedding data which B can possibly make sense of. This would, on the one hand, undermine arguments in favor of B over A on the basis of embedding data. And, provided that the converse does not hold—that is, that A can make sense of strictly more embedding data than B can—it would also show that there is a precise sense in which B is more constrained than A, yielding a *pro tanto* simplicity-based consideration in favor of B. In this paper I develop tools which allow us to make comparisons of this kind, which I call comparisons of *potential expressive power*. I motivate the development of these tools by way of exploration of the recent debate about epistemic modals. Prominent theories which have been developed in response to embedding data turn out to be strictly less expressive than the standard relational theory, a fact which necessitates a reorientation in how to think about the choice between these theories.

Keywords: relative expressibility; semantic theories; epistemic modals; attitude predicates

1 Introduction

When embedding data are used to argue against semantic theory A and in favor of semantic theory B, it is important to ask whether A could, after all, make sense of those data. It is possible to ask that question on a case-by-case basis. But suppose we could show that A can make sense of *all* the embedding data which B can make sense of. This would, on the one hand, undermine arguments in

*This paper is a descendant of work presented at MIT, the Eighth Philosophy and Semantics in Europe Colloquium at the University of Cambridge, the 2015 Bucharest Colloquium for Analytic Philosophy, and the 2016 Central APA, and of the final chapter of Mandelkern 2017a. Thanks to audiences at those talks and to Justin Bledin, David Boylan, Daniel Drucker, Kai von Fintel, Vera Flocke, Peter Fritz, Cosmo Grant, Irene Heim, Valentine Hacquard, Matthias Jenny, Roni Katzir, Justin Khoo, Angelika Kratzer, Harvey Lederman, Jack Marley-Payne, Vann McGee, Maša Močnik, Dilip Ninan, Milo Phillips-Brown, Agustín Rayo, Daniel Rothschild, Bernhard Salow, Paolo Santorio, Ginger Schultheis, Roger White, and Stephen Yablo for very helpful comments and discussion on these topics. Thanks to Fabrizio Cariani, Simon Goldstein, Jonathan Phillips, Robert Stalnaker, Shane Steinert-Threlkeld, and three anonymous referees for this journal for extensive comments and discussion.

†All Souls College, Oxford, OX1 4AL, United Kingdom, matthew.mandelkern@all-souls.ox.ac.uk

favor of B over A on the basis of embedding data. And, provided that the converse does not hold—that is, that A can make sense of strictly more embedding data than B can—it would also show that there is a precise sense in which B is more constrained than A, yielding a *pro tanto* simplicity-based consideration in favor of B.

In this paper I develop formal tools which allow us to make comparisons of this kind, which I call comparisons of *potential expressive power*. I motivate the development of these tools through discussion of the recent debate about epistemic modals (words like 'might' and 'must', on a broadly epistemic reading). Recent work has used embedding data to argue against the standard relational theory—on which epistemic modals have, roughly, the semantics of modal operators in standard modal logics—and in favor of various revisionary theories. But comparisons of potential expressive power show that those revisionary theories are strictly less expressive than the standard theory, in the sense that the relational theory can make sense of any embedding data involving epistemic modals which those theories can make sense of, but not *vice versa* (within very mild limits). This necessitates a reorientation in how to think about the choice between these semantics. The situation is roughly the opposite of the way it is standardly presented: we can rest assured that the relational theory can account for any data which those revisionary theories can account for. But not *vice versa*: we may well discover data that the relational theory can make sense of, but that the revisionary theories cannot. This, in turn, shows that there is a precise sense in which these revisionary theories are simpler than the relational theory, which may count in their favor.

The tools developed here allow us to state these facts in a rigorous and general way, which, among other things, makes application to other debates straightforward. Considerations of relative potential expressive power cannot on their own decide between two semantic theories, but they can help us determine what *kinds* of arguments from natural language on the basis of embedding data are possibly good arguments in the choice between those theories. In the final part of the paper, I consider in more detail the empirical picture concerning the behavior of epistemic modals in attitude contexts, and what this behavior tells us about the theory of epistemic modals in light of the foregoing expressibility results.

2 Background

I begin by introducing the relational semantics for epistemic modals; Yalcin (2007)'s data and his revisionary *domain semantics* response to those data; and Ninan (2016)'s response to Yalcin, which shows that the relational semantics can replicate the predictions of the domain semantics.

2.1 The relational theory

On the relational theory (e.g. Kripke 1963; Kratzer 1977, 1981), epistemic modals denote quantifiers over a set of worlds: namely, over those worlds which are accessible from the world of evaluation relative to a binary *accessibility relation* between worlds (equivalently and more conveniently, a corresponding *modal base* function from worlds to sets of worlds). The accessibility relation for epistemic modals is generally taken to track the contextually relevant evidence: a world is accessible from another world just in case it's compatible with the contextually relevant evidence in that world. 'Might' is treated as an existential quantifier, so ⌜Might φ⌝ is a claim that the contextually relevant evidence leaves it open that φ is true;[1] 'must' is treated as the dual universal quantifier, so ⌜Must φ⌝ is a claim that the contextually relevant evidence entails that φ is true. More formally:[2]

Definition 2.1. *Relational Semantics:* Where f is a modal base:

- $[\![\text{Might } \varphi]\!]^{f,w} = 1$ iff $\exists w' \in f(w) : [\![\varphi]\!]^{f,w'} = 1$
- $[\![\text{Must } \varphi]\!]^{f,w} = 1$ iff $\forall w' \in f(w) : [\![\varphi]\!]^{f,w'} = 1$

2.2 Yalcin (2007)'s challenge

Yalcin (2007) observed that, when you embed a sentence with the form ⌜φ and might not φ⌝ (which he calls an *epistemic contradiction*) under 'Suppose', the result is infelicitous:[3]

(1) a. #Suppose it's raining and it might not be raining.
b. #Suppose (φ and might not φ).

The issue for the relational approach is that it seems to predict that a sentence like (1-a) will be perfectly felicitous. To spell out the problem, let's focus (as Yalcin does) on the corresponding declarative sentence. A sentence like (2) is felt to attribute inconsistent suppositions to Alfred, a fact which explains why the imperative variant will be felt to be infelicitous:

(2) a. Alfred supposes it's raining and it might not be raining.

[1] Greek letters range over all sentences. Italic roman letters below range over atomic sentences. I will leave off a context parameter throughout, for readability; insofar as the languages we are working with contain context-sensitive terms (beyond epistemic modals), these should be read as implicit throughout.

[2] This presentation simplifies Kratzer's theory in two ways, one irrelevant (for Kratzer, a modal base is a function from worlds to sets of propositions, not worlds, and quantification is over the intersection of that set), and one possibly relevant (for Kratzer, modals are evaluated relative to a second parameter, an *ordering source*). Insofar as an ordering source plays in interesting role in what follows, though, it makes the relational theory *even more expressive* than it is on the present simplification.

[3] Parallel observations concerning the embedding of sentences with this form in the scope of quantifiers were made in Groenendijk et al. 1996; Aloni 2001, and were used to motivate the update semantics, discussed below. I focus on Yalcin's version of the argument only because it is simpler.

b. A supposes (φ and might not φ).

But, given the relational semantics, it looks like (2-a) is predicted to say that Alfred supposes that it's raining, and supposes that the contextually salient body of evidence is compatible with the proposition that it's not raining. In other words, (2-a) should be roughly equivalent to (3), which is not felt to attribute any kind of inconsistency to Alfred.

(3) Alfred supposes that it's raining and that, for all he knows, it's not raining.

To make this worry more precise, we need to spell out some background assumptions which, as we will see, are crucial for getting the puzzle going. First, Yalcin assumes that connectives have classical Boolean semantics.[4] Second, Yalcin generalizes Hintikka (1962)'s theory of attitude predicates to the relational semantics as follows. On Hintikka's approach, attitude predicates denote universal quantifiers over the possible worlds compatible with the relevant attitude (for attitude verb V and agent A, let '$V_{A,w}$' denote the worlds compatible with everything that A V's in w). Yalcin assumes further that attitude verbs do not change the setting of the modal base parameter, so the schematic result is as follows:

Definition 2.2. *Simple Hintikkan semantics*:

- $[\![\text{A V's } \varphi]\!]^{f,w} = 1$ iff $\forall w' \in V_{A,w} : [\![\varphi]\!]^{f,w'} = 1$

Thus, in particular, $\ulcorner$A supposes $\varphi\urcorner$, evaluated at modal base f and world w, just says that φ is true at every world compatible with A's suppositions in w, relative to f. Together with the relational semantics for epistemic modals and Boolean semantics for connectives, we thus predict that $\ulcorner$A supposes (φ and might not φ)$\urcorner$, as evaluated at modal base f and world w, says that every world in A's supposition state makes φ true (relative to f), and that every world in A's supposition state can access under f some world where φ is false. So, if the modal base tracks something like A's knowledge, then (2-a) should just mean something like (3), and thus should not be felt to ascribe incompatible suppositions to Alfred.

2.3 Yalcin (2007)'s domain semantics

This is a puzzle for the combination of the relational semantics with the simple Hintikkan semantics for attitude verbs given here together with the Boolean semantics for connectives. Yalcin (2007) responded to this puzzle by developing a theory that is revisionary twice over, rejecting both the relational theory of epistemic modals and the simple Hintikka semantics for attitude verbs. First, Yalcin

[4] That is, $[\![\varphi \text{ and } \psi]\!]^{f,w} = 1$ iff $[\![\varphi]\!]^{f,w} = 1$ and $[\![\psi]\!]^{f,w} = 1$; and $[\![\text{Not } \varphi]\!]^{f,w} = 1$ iff $[\![\varphi]\!]^{f,w} = 0$.

adopts a *domain semantics* for epistemic modals (crediting an early version of MacFarlane 2011). On the domain semantics, epistemic modals denote quantifiers over a set of worlds (*information state*) s which is supplied, not by an accessibility relation, but rather as a world-independent parameter of the index.

Definition 2.3. *Domain semantics*:

- $[\![\text{Might } \varphi]\!]^{s,w} = 1$ iff $\exists w' \in s : [\![\varphi]\!]^{s,w'} = 1$
- $[\![\text{Must } \varphi]\!]^{s,w} = 1$ iff $\forall w' \in s : [\![\varphi]\!]^{s,w'} = 1$

Thus ⌜Might φ⌝ says that the truth of φ is compatible with a given information state; ⌜Must φ⌝ says that the truth of φ is entailed by a given information state. The domain semantics is just like the relational semantics except that it substitutes information states where the relational semantics has functions from worlds to information states. Let me note that I will use 'domain semantics' and 'relational semantics' to denote just the semantics for epistemic modals given here and above, separate from the semantics for attitude verbs and connectives that each of these theories has been typically coupled with (I use 'relational/domain framework' for a semantic theory which extends the relational/domain theories). This lets us focus on the question of how those particular semantics for epistemic modals can be motivated on their own, rather than as part of a package deal.

Yalcin generalizes the classical semantics for the connectives to the domain semantics in the obvious way.[5] Yalcin, finally, develops a new semantics for attitude verbs, which builds on the core Hintikka semantics, but stipulates that the attitude verb supplies the set of attitude worlds as the domain of quantification for its complement. Schematically, for attitude verb V, Yalcin proposes:[6]

Definition 2.4. *Yalcin semantics for attitude verbs*:

- $[\![\text{A V's } \varphi]\!]^{s,w} = 1$ iff $\forall w' \in V_{A,w} : [\![\varphi]\!]^{V_{A,w},w'} = 1$.

That is, ⌜A V's φ⌝ is true at w and s just in case φ is true at every world compatible with what A V's at w, *relative to* $V_{A,w}$.

This combination of views accounts for Yalcin's data as follows. Consider a sentence with the form of (4):

(4) A supposes (φ and might not φ).

Given Yalcin's approach to attitude verbs (plus Boolean connectives), (4) will be true at w, relative to any information state, just in case, at every world w' in $S_{A,w}$ (the set of worlds compatible with A's

[5] Namely, $[\![\varphi \text{ and } \psi]\!]^{s,w} = 1$ iff $[\![\varphi]\!]^{s,w} = 1$ and $[\![\psi]\!]^{s,w} = 1$, and $[\![\text{Not } \varphi]\!]^{s,w} = 1$ iff $[\![\varphi]\!]^{s,w} = 0$.

[6] Yalcin presents a semantics only for 'supposes', but suggests that it should be extended to other attitudes in the obvious way; for simplicity, I am presenting the general format here. Likewise for Ninan's semantics below.

suppositions in w), $\ulcorner\varphi$ and might not $\varphi\urcorner$ is true relative to $\langle S_{A,w}, w'\rangle$. This follows just in case, for every world w' in $S_{A,w}$, φ is true at $\langle S_{A,w}, w'\rangle$; and, for every world w' in $S_{A,w}$, there is some world w'' in $S_{A,w}$ such that φ is false at $\langle S_{A,w}, w''\rangle$. Clearly these two conditions can only both be met if $S_{A,w}$ is empty. Thus a sentence with the form of (4) will ascribe inconsistent suppositions to A. This, again, suffices to account for Yalcin's data: the imperative version of (4) will be infelicitous because it will be a command to make an inconsistent supposition.[7]

2.4 Ninan (2016)'s response

Yalcin's proposal is, again, revisionary in two respects: it couples a new semantics for modals (the domain semantics) with a new semantics for attitude verbs. Ninan (2016) shows that there is a way to account for Yalcin's data while maintaining the relational semantics for epistemic modals. For any set of worlds s, let f^s be the constant function which takes any world to s. Then, for attitude verb V, modal base f, and world w, Ninan proposes:

Definition 2.5. *Ninan semantics for attitude verbs*:

- $[\![\text{A V's } \varphi]\!]^{f,w} = 1$ iff $\forall w' \in V_{A,w} : [\![\varphi]\!]^{f^{V_{A,w}},w'} = 1$.

The resemblance to Yalcin's attitude semantics is clear, and this approach accounts for Yalcin's data in essentially the same way that Yalcin's semantics does, but within a relational framework. To see this, suppose $\ulcorner$A supposes (φ and might not φ)$\urcorner$ is true at world w and modal base f. Given Ninan attitude semantics, plus the relational semantics for modals and Boolean connectives, this holds just in case, for every world w' compatible with A's suppositions at w, w' has the following properties: (i) φ is true at $\left\langle f^{S_{A,w}}, w'\right\rangle$; (ii) there is some world w'' in $f^{S_{A,w}}(w')$ such that φ is false at $\left\langle f^{S_{A,w}}, w''\right\rangle$. Given that $f^{S_{A,w}}(w') = S_{A,w}$, it is easy to confirm that these conditions are only met when A's suppositions are inconsistent, and so $\ulcorner$A supposes (φ and might not φ)$\urcorner$ will entail that A's suppositions are inconsistent, as in Yalcin's system.

3 Relative potential expressibility

Is it a fluke that Ninan was able to reproduce, in a relational framework, the interaction of Yalcin's attitude verbs with the domain semantics—or is there a deeper fact about the relationship between the domain and relational semantics which accounts for this? This is an important question for both practical and theoretical reasons. Practically, this is an important question because Yalcin's data are not

[7] Yalcin (2007) in fact claims that, on his semantics, a sentence with the form $\ulcorner$A supposes (φ and might not φ)$\urcorner$ is itself a contradiction, but this would only follow if we add a presupposition that the attitude is consistent, as Ninan (2016) proposes.

the only challenge to the relational semantics from embedding behavior. Yalcin 2007 discusses similar data involving epistemic contradictions in the antecedents of conditionals, and related challenges on epistemic modals in a variety of different embedding environments has raised related challenges for relational approaches and been used to argue in favor of revisionary semantics.[8] Ninan's system shows that the relational semantics can make sense of just some of these data. One way to proceed would be to go through each data point, and see whether the relational framework can replicate the predictions of the revisionary accounts with respect to those data. But this approach is inefficient; and, more problematically, even if this method showed that all known data can be accounted for in the relational framework, this would still leave open the possibility that more data are lurking undiscovered which a revisionary approach can account for, but which the relational semantics cannot account for. It would be far more helpful if we could show that the relational framework can account for *any* embedding data which can be accounted for in the revisionary frameworks. From a more theoretical point of view, answering this question will elucidate the structure of the underlying relationship between the relational semantics and its competitors. And intuitively, it is very natural to think that there is a sense in which the domain semantics is simpler and more constrained than the relational semantics, as Ninan (2016) points out. In this section I will spell out a general framework which allows us to spell out these intuitions in a precise way, by comparing the expressive power of two theories of a given fragment—like the relational and domain semantics—with respect to what embedding operators are definable within those two theories.

Let me first note that, in the choice between semantic theories, embedding data invariably provide just one motivation. There will also be framework-level considerations which are at least somewhat independent of particular embedding data. In the debate about epistemic modals in particular, issues about assertability, disagreement, and retraction have played a crucial role. Thus, for instance, Yalcin advocates the domain semantics as part of an *expressivist* approach to epistemic modality; MacFarlane advocates it as part of a *relativist* approach. Ninan, for his part, remains neutral about the 'post-semantics' he intends for his semantics, but the relational semantics is most commonly coupled with a *contextualist* post-semantics (though it needn't be). The choice between these frameworks turns, at least in substantial part, on issues that are independent of embedding data.[9] As Ninan (2010) discusses in detail, while it is possible that embedding data will have some impact on the choice between these frameworks, the impact will be at best indirect; embedding data tell us about the logic of the operator we are studying, but not about assertion, disagreement, truth, and so on.[10] I will focus in this paper

[8] See Beaver 1994; Groenendijk et al. 1996; Gerbrandy 1998; Aloni 2000, 2001; Yalcin 2015; Rothschild and Klinedinst 2015; Ninan 2018; Mandelkern 2019.

[9] See e.g. Lewis 1980; Ninan 2010; Rabern 2012, 2013 on the distinction between assertoric and semantic content; for relativist approaches, see e.g. Egan et al. 2005; Stephenson 2007; MacFarlane 2011, 2014; for expressivism, e.g. Yalcin 2007; Swanson 2015; Moss 2015; for contextualism, e.g. Dowell (2011); Khoo (2015); Mandelkern (2018a).

[10] Though there may be an indirect bearing—for instance, the logical facts may go more naturally with one conception

narrowly on the choice of semantic theories, not post-semantic theories.

Abstracting from the particulars of epistemic modals, the kind of question that I want to address is the following. Given two semantic theories of a given language, suppose that we add an arbitrary new sentence operator to the language, and extend the first semantic theory to give a compositional semantics for the resulting language.[11] Can we be *guaranteed* to find a way of extending the *second* semantic theory to give a compositional semantics for that operator such that the operator has *exactly the same logic*, according to the second theory, as it does according to the first theory? If so, we know that the first theory is no more expressive than the second theory with respect to embedding data. That is, for any embedding data involving an operator O not covered by the first theory, if the first theory can be extended to make sense of those data—that is, if we can extend the first theory so that we have a semantics for O which makes all the right predictions about O's logic—then we can also cover the same data in the second theory: we can give a semantics for O in the second theory which makes all the right predictions about O's logic.

We can spell this out more precisely as follows. First, we make precise the notion of a semantic theory (which I'll denote $\mathcal{T}, \mathcal{T}'$, and so on) for a language (a set of sentences $\mathcal{L}$) which is meant to correspond roughly to the kinds of systems that semanticists construct to make sense of fragments of natural language. For propositional languages, we will let our *models* be sequences comprising a set of possible worlds; a valuation function which takes atomic sentences in the language to subsets of the set of possible worlds; a set of indices; and an interpretation function. Indices ordinarily will be sequences which include a possible world or set of possible worlds (from the stock of worlds in the model), and may also include other parameters, like an accessibility relation or set of worlds. Interpretation functions interpret the language relative to the set of indices, given the model's valuation function: for any sentence and index, the interpretation function tells us whether the sentence is true or false at that index. This extended notion of a model (relative to a standard account in modal logic) allows us to schematize in a general way very different kinds of semantic theories, and thus to compare those approaches. Finally, a *semantic theory* of a language is any set of models of that language. A semantic theory for a language yields a logic for the language: for a set of sentences Γ and sentence ψ from the language, $\Gamma \vDash_{\mathcal{T}} \psi$ means that, in every model in $\mathcal{T}$, ψ is true at every index where all the elements of Γ are true.[12] All of this is spelled out more formally in Appendix A.

With this in hand, we can define our notion of *relative potential expressibility* as follows:

Definition 3.1. *Relative Potential Expressibility*: Given two semantic theories $\mathcal{T}$ and $\mathcal{T}'$ for a language

of logical consequence than another, as e.g. Veltman (1996); Groenendijk et al. (1996); Yalcin (2007) have argued with regards to epistemic modals.

[11] By 'compositional', I mean that the meaning of a sentence containing the operator in question with scope over some sequence of sentences is obtained by applying the meaning of the operator (which we assume is an n-place propositional function) to the meaning of the sentences it takes scope over.

[12] I write $\varphi \vDash_{\mathcal{T}} \psi$ for $\{\varphi\} \vDash_{\mathcal{T}} \psi$.

$\mathcal{L}, \mathcal{T}$ is no more expressive than $\mathcal{T}'$ with respect to $\mathcal{L}$ (written $\mathcal{T} \preceq_{\mathcal{L}} \mathcal{T}'$) iff, for any set of new sentence operators $\mathcal{O}$, for any extension $\mathcal{T}^{\mathcal{O}}$ of $\mathcal{T}$ to $\mathcal{L}^{\mathcal{O}}$, there is an extension $\mathcal{T}'^{\mathcal{O}}$ of $\mathcal{T}'$ to $\mathcal{L}^{\mathcal{O}}$ which preserves the logic of $\mathcal{O}$ from $\mathcal{T}^{\mathcal{O}}$.

$\mathcal{L}^{\mathcal{O}}$ is the closure of $\mathcal{L}$ under the elements of $\mathcal{O}$. What it means for $\mathcal{T}'^{\mathcal{O}}$ to preserve the logic of $\mathcal{O}$ from $\mathcal{T}^{\mathcal{O}}$ is the following: for any set of sentences $\Gamma \subseteq \mathcal{L}^{\mathcal{O}}$ at least one of which contains an operator from $\mathcal{O}$ (i.e. one of which is in $\mathcal{L}^{\mathcal{O}} \setminus \mathcal{L}$), for any sentence ψ in $\mathcal{L}^{\mathcal{O}}$, $(\Gamma \vDash_{\mathcal{T}^{\mathcal{O}}} \psi) \leftrightarrow (\Gamma \vDash_{\mathcal{T}'^{\mathcal{O}}} \psi)$. (More details, again, in the appendix; '$\leftrightarrow$' abbreviates meta-language 'iff'.)

The notion of relative potential expressibility spelled out here is somewhat different from the extant notions of expressibility I know of, which are typically concerned with comparisons between *different languages* rather than with comparisons between different *semantic theories* of a language with respect to *arbitrary extensions of the language*. But relative potential expressibility is just what we need to answer the questions posed at the beginning of this section. If $\mathcal{T} \preceq_{\mathcal{L}} \mathcal{T}'$, this tells us that $\mathcal{T}'$ can do anything $\mathcal{T}$ can with regard to arbitrary extensions of $\mathcal{L}$, and thus that $\mathcal{T}'$ can make sense of any embedding data that $\mathcal{T}$ can make sense of.[13]

Before exploring the application of this framework, let me note a helpful characterization result. Relative potential expressibility between certain kinds of well-behaved semantic theories boils down to something very simple: the existence of a truth-preserving injection from the indices of a model in the first theory to those of a model in the second.

Fact 3.1. *Characterization of Expressibility*: For any semantic theories $\mathcal{T}$ and $\mathcal{T}'$ and language $\mathcal{L}$, if $\mathcal{T}$ is isomorphic with respect to $\mathcal{L}$, and $\mathcal{T}'$ is isomorphic with respect to $\mathcal{L}$, then $\mathcal{T} \preceq_{\mathcal{L}} \mathcal{T}'$ iff there is a model $\mathcal{M} \in \mathcal{T}$ and a model $\mathcal{M}' \in \mathcal{T}'$ such that there is a injection g from the indices of $\mathcal{M}$ to those of $\mathcal{M}'$ which preserves the truth of all sentences in $\mathcal{L}$, i.e. such that $\forall \varphi \in \mathcal{L}$, for any index i from $\mathcal{M}$, φ is true at i in $\mathcal{M}$ iff φ is true at $g(i)$ at $\mathcal{M}'$. A semantic theory $\mathcal{T}$ is *isomorphic* with respect to $\mathcal{L}$ iff $\forall \mathcal{M}, \mathcal{M}' \in \mathcal{T} : \mathcal{M} \preceq_{\mathcal{L}} \mathcal{M}' \wedge \mathcal{M}' \preceq_{\mathcal{L}} \mathcal{M}$.

The proof of Fact 3.1, and of all the other facts stated in the rest of this section, are relegated to Appendix A. I state this characterization result here because it greatly simplifies the proofs of the facts that follow, and, I think, provides some intuitive grip on those results: whether one semantic theory is less expressive than another (in the relevant sense) depends on the structural relations between the semantic theories, and in particular whether the first semantic theory can, from a very abstract perspective, be embedded into the second semantic theory in a truth-preserving way.

[13] There is some work on comparing expressive power across different classes of structures, though not to my knowledge anything along the lines of what I am proposing here; see Pinheiro Fernandes 2017 for an overview and relevant citations, in particular Mossakowski et al. 2009, which proposes a structurally similar framework (in particular in their definition of sublogics), but one that remains quite different in its details. Cf. also French (2017) for investigation of the notion of notational equivalence over possible extensions of the language.

Let me make a final note about the relation between relative potential expressibility and the relative strength of logics. If $\mathcal{T} \preceq_{\mathcal{L}} \mathcal{T}'$, then the logic of $\mathcal{L}$ in $\mathcal{T}'$ can be no stronger than the logic of $\mathcal{L}$ in $\mathcal{T}$ (they may have the same logic, but it may also be that the logic of $\mathcal{L}$ in $\mathcal{T}$ is strictly stronger than the logic of $\mathcal{L}$ in $\mathcal{T}'$).[14] But it is not guaranteed to be weaker (even if $\mathcal{T}$ is strictly less expressive than $\mathcal{T}'$, i.e. $\mathcal{T} \prec_{\mathcal{L}} \mathcal{T}'$). Nor does the converse follow: if the logic of $\mathcal{L}$ in $\mathcal{T}'$ is no stronger than the logic of $\mathcal{L}$ in $\mathcal{T}$, there is no guarantee that $\mathcal{T} \preceq_{\mathcal{L}} \mathcal{T}'$; likewise, if the logic of $\mathcal{L}$ in $\mathcal{T}'$ is strictly weaker than the logic of $\mathcal{L}$ in $\mathcal{T}$, there is no guarantee that $\mathcal{T} \preceq_{\mathcal{L}} \mathcal{T}'$. So, while relative potential expressibility has some bearing on relative logical strength, these two ways of comparing semantic theories are orthogonal. (Even if they turned out to coincide, this would itself be a substantive fact, since it would not have been at all obvious that they should coincide; that they fail to coincide shows that they get at distinct features of semantic theories).[15]

3.1 The domain and relational semantics

With this discussion in hand, I will turn my attention to the specific comparisons of expressive power which I will focus on here: namely, comparisons between different semantics for epistemic modals. A note to readers: the rest of this section is concerned with introducing a variety of semantics for epistemic modals and comparing their expressive power. I succinctly summarize these results at the beginning of the next section (§4); readers uninterested in the details that follow may wish to skim up to that point.

Since our focus is on epistemic modals, I will compare semantic theories for a very simple language $\mathcal{L}^{\Diamond}$ comprising infinitely many atomic sentences $p_i : i \in \mathbb{N}$ (which I will write p, q, r etc.) together with sentences of the form $\Diamond\varphi$, for any sentence $\varphi \in \mathcal{L}^{\Diamond}$ ($\Diamond$ abbreviates 'might').[16] Let $\mathcal{R}$ denote the relational semantics, and $\mathcal{D}$ the domain semantics. These are, again, classes of models, in

[14] It may look as though, if $\mathcal{T} \preceq_{\mathcal{L}} \mathcal{T}'$, then we can prove that the logic of $\mathcal{L}$ in $\mathcal{T}'$ must be the same as in $\mathcal{T}$: we simply introduce a one-place sentence operator *Id* to $\mathcal{L}$, and give *Id* the semantics in $\mathcal{T}$ of the identity function: for all φ, $[\![Id(\varphi)]\!]_{\mathcal{T}} = [\![\varphi]\!]_{\mathcal{T}}$. Then we will be guaranteed to be able to give *Id* a semantics in $\mathcal{T}'$ such that the logic of *Id* in the extension of $\mathcal{T}'$ matches that of *Id* in the extension of $\mathcal{T}$. But the fact that we can do this does *not*, contrary to first appearances, guarantee that $\mathcal{L}$ has the same logic in the extension of $\mathcal{T}'$ as in the extension of $\mathcal{T}$, because nothing guarantees that *Id* will still denote the identity function in the extension of $\mathcal{T}'$—this is no part of the logic of *Id*.

[15] An anonymous referee for this journal points out that there are, however, interesting cases in which these two notions do coincide. It seems to me both a striking fact that they do not coincide in general, and that there are limited cases where they do, and it seems worth further exploration exactly when they coincide, and why. Things would have been otherwise had we defined $\mathcal{T} \preceq_{\mathcal{L}} \mathcal{T}'$ in a slightly different way, such that this means that for arbitrary extension $\mathcal{L}^{\mathcal{O}}$ of $\mathcal{L}$ and extension $\mathcal{T}^{\mathcal{O}}$ of $\mathcal{T}$, we can extend $\mathcal{T}'$ to a model $\mathcal{T}'^{\mathcal{O}}$ which matches the logic of $\mathcal{L}^{\mathcal{O}}$ in $\mathcal{T}^{\mathcal{O}}$, i.e. which is such that for any $\Gamma \cup \{\psi\} \subseteq \mathcal{L}^{\mathcal{O}}$, $(\Gamma \vDash_{\mathcal{T}^{\mathcal{O}}} \psi) \leftrightarrow (\Gamma \vDash_{\mathcal{T}'^{\mathcal{O}}} \psi)$. We instead said that $\mathcal{T}'^{\mathcal{O}}$ only must match the logic of $\mathcal{O}$ in $\mathcal{T}^{\mathcal{O}}$. This means that $\mathcal{T} \preceq_{\mathcal{L}} \mathcal{T}'$ does not at all entail that $\mathcal{T}$ and $\mathcal{T}'$ have to agree on the logic of $\mathcal{L}$. This, in turn, makes it possible to compare the relative expressive power of semantic theories which start with different logics for a given language—as the domain and relational semantics do—with respect to possible *enrichments* of those languages, something which, as we will see, turns out to be very revealing. Otherwise comparisons of relative potential expressive power would be limited to starting languages with identical logics.

[16] I let sentences of the language name themselves for brevity.

the sense given above; I assume that the models in every case contain sets of possible worlds and valuation functions such that for any set of atoms, exactly that set is true at some world, and any two worlds differ on the truth of some atom. Recall that for any model $\mathfrak{r} \in \mathcal{R}$, an index in that model is any pair $\langle f, w \rangle$, with f a modal base and w a possible world drawn from $\mathfrak{r}$'s stock of worlds. An atomic sentence p is true in $\mathfrak{r}$ at $\langle f, w \rangle$ just in case the model's valuation function takes p to true at w; and $\Diamond\varphi$ is true in $\mathfrak{r}$ at $\langle f, w \rangle$ just in case there is a world w' in $f(w)$ such that φ is true at $\langle f, w' \rangle$ according to $\mathfrak{r}$. Recall that an index in any model $\mathfrak{d} \in \mathcal{D}$ is a pair $\langle s, w \rangle$, with s an information state and w a possible world. An atomic sentence p is true in $\mathfrak{d}$ at $\langle s, w \rangle$ just in case p is true at w according to $\mathfrak{d}$'s valuation function. A sentence with the form $\Diamond\varphi$ is true at $\langle s, w \rangle$ in $\mathfrak{d}$ just in case there is a world $w' \in s$ such that φ is true at $\langle s, w' \rangle$ according to $\mathfrak{d}$.

We have some flexibility in how we think about entailment in the domain framework. The classical notion of entailment would say that for any domain model $\mathfrak{d}$, $\Gamma \models_{\mathfrak{d}} \psi$ iff ψ is true at every index in $\mathfrak{d}$ where all the members of Γ are. But a different route, suggested by Yalcin (2007), would be to treat entailment as preservation of *acceptance* rather than truth, where an index $\langle s, w \rangle$ accepts a sentence φ iff for every world $w' \in s$, φ is true at $\langle s, w' \rangle$. Preservation of truth is a strictly stronger notion than preservation of acceptance: if ψ is true at every index where all the members of Γ are, then ψ is also accepted at every index where all the members of Γ are. But preservation of acceptance does not entail preservation of truth; to take a simple example, p entails $\Diamond p$ in the preservation of acceptance sense, but not in the preservation of truth sense (at least not without further stipulation about our models).[17] So which logical notion will be the most illuminating for our purposes? I think it is the preservation of truth notion, for a few reasons. For one thing, on the preservation of acceptance notion, the domain semantics ends up being equivalent to the state-based semantics discussed below. So, just for the sake of diversity, it is worth exploring a different perspective on the domain semantics; readers who are interested in what we would find with a preservation of acceptance notion are referred to the discussion of the state-based semantics below. Second, this perspective renders the domain semantics *more expressive* than it otherwise would be; the acceptance-based approach renders the world parameter of indices essentially invisible, bleaching the semantics of a good deal of expressive strength (as we will see below in the discussion of the state-based semantics). There are many ways to think about the logic of a given system; I think that for present purposes it makes sense to take a vantage point which does not wash out any part of the semantics' inherent expressive power. This allows us to explore the expressive power of the semantics itself, as it were, rather than of the semantics under a certain, possibly limiting, perspective. Thus I will stick with the preservation of truth notion of consequence for both the relational and domain semantics.

[17] Yalcin (2007) countenances the possibility that a domain index $\langle s, w \rangle$ should always be such that $w \in s$, in which case this disanalogy would drop out. With or without this addition, the acceptance-based logic remains strictly weaker than the truth-based logic across extensions of the language.

With this in mind, we can show that the domain semantics is no more expressive than the relational semantics:

Fact 3.2. $\mathcal{D} \preceq_{\mathcal{L}^\Diamond} \mathcal{R}$.

Proofs of all these results are, again, included in the appendix, but I will say a bit here about the intuition behind each proof. In this case, the proof strategy is to choose an arbitrary model in $\mathcal{D}$ and construct a function g which takes any index $\langle s, w\rangle$ in that model to $\langle f^s, w\rangle$, where f^s is the constant function from worlds to s. Then Fact 3.2 follows from our characterization result in Fact 3.1.[18]

Next we show that the relational semantics is not less expressive than the domain semantics with respect to $\mathcal{L}^\Diamond$:

Fact 3.3. $\mathcal{R} \not\preceq_{\mathcal{L}^\Diamond} \mathcal{D}$.

Fact 3.3 follows immediately from the observation that, for any index i in any model in $\mathcal{D}$, $\Diamond\varphi$ is true$_\mathcal{D}$ at i iff $\Diamond(\Diamond\varphi)$ is; but the same does not hold in $\mathcal{R}$. This will make it impossible to find a truth-preserving injection from the indices of any $\mathcal{R}$ to those of any $\mathcal{D}$ model. I should note here that this difference in logics *in the logic of* $\mathcal{L}^\Diamond$ *itself* may be taken on its own as an argument for, or against, the domain semantics, depending on what one thinks of the claim that $\Diamond(\Diamond\varphi)$ is always equivalent to $\Diamond\varphi$—a claim, again, validated by $\mathcal{D}$ but not by $\mathcal{R}$. Yalcin claims this kind of logical feature as a virtue of his theory, while Moss (2015) criticizes this prediction. I will not explore that debate here, though. Our concern is with a very different question: ignoring the logical differences between the domain and relational semantics within $\mathcal{L}^\Diamond$, which theory is better equipped to account for embedding data which involve operators beyond our starting language $\mathcal{L}^\Diamond$? It is this latter question I focus on here, though the former may, of course, play an important independent role in the final choice between these theories.

Thus from Facts 3.2 and 3.3, we have that the domain semantics is strictly less expressive than the relational semantics with respect to $\mathcal{L}^\Diamond$.

Fact 3.4. $\mathcal{D} \prec_{\mathcal{L}^\Diamond} \mathcal{R}$.

It is worth noting that 'might's are not generally able to stack in natural language: that is, sentences with the form ⌜Might (Might φ)⌝ are generally infelicitous.[19] This may make it seem as though Fact 3.3 holds due to an irrelevant technicality. But this is not so: we still have an expressive inequality between the domain and relational semantics even if we add a syntactic constraint which rules out embedded 'might's (and thus which renders the logics of each theory for this more limited language fully equivalent). In other words, where $\mathcal{L}^{\Diamond-}$ is the language comprising just atomic sentences from $\mathcal{L}^\Diamond$ and sentences of the form $\Diamond p_i$, where p_i is any atom, we still have:

[18] Note that it is easy to extend this to a proof that the domain semantics is no more expressive than the relational semantics with respect to an expanded language containing not just atoms and 'might' sentences, but also sentences formed with the Boolean connectives.

[19] Though see Moss 2015 for discussion of nested epistemic modals.

Fact 3.5. $\mathcal{D} \prec_{\mathcal{L}^{\Diamond-}} \mathcal{R}$.

The intuition behind these facts is that sets of worlds are in some sense less fine-grained than functions from worlds to sets of worlds (though this is not a fact about cardinality: given a fixed infinite stock of worlds, the set of domain indices and the set of relational indices have the same cardinality).

3.2 The update and state-based semantics

An important fact about the formal tools developed here is that it is straightforward to apply them to further comparisons: nothing in the kind of comparison we are doing here depends on details of the domain or relational semantics. In the rest of this section, I will illustrate this by exploring the relative potential expressive power of a few more revisionary semantics of epistemic modals which have been motivated by embedding data. In each case, the relational semantics comes out as the most expressive option.

First, the comparison of expressive power can be extended to the update semantics. In the update semantics $\mathcal{U}$, due to Veltman (1996), building on Heim 1982, 1983, the intension of a complex sentence is a dynamic object: a function from information states (usually called *contexts* in this literature) to information states. The most natural way to think about this approach from the point of view of the formal framework we are using is to think of our "indices" as pairs of contexts, with φ "true" at a pair $\langle s, c\rangle$ in $\mathcal{U}$ just in case $\langle s, c\rangle \in [\![\varphi]\!]_{\mathcal{U}}$. As for the domain semantics, we have different options for thinking about entailment. The option most parallel to the approach we've taken so far would treat entailment as preservation of truth, so that $\varphi \models_{\mathcal{U}} \psi$ just in case the function denoted by φ in $\mathcal{U}$ is a subset of a function denoted by ψ in $\mathcal{U}$ (likewise, *ceteris paribus*, for multi-premise entailment). This is a non-standard way of thinking about entailment in the dynamic framework. A different approach would be to treat our indices simply as contexts, rather than pairs of contexts, and then treat entailment as preservation of *acceptance*, where a context c accepts a sentence φ just in case c is a fixed point for φ, i.e. just in case $[\![\varphi]\!]_{\mathcal{U}}(c) = c$. This would bring our notion of entailment in line with Veltman (1996)'s (third) notion of entailment as preservation of acceptance. Just as for the domain semantics, however, I think the truth-based notion is the more interesting one, for essentially the same reasons. First, on an acceptance-based notion, the update semantics ends up, again, equivalent for our purposes to the state-based semantics, so the discussion of the state-based semantics below will suffice to show what would result from an acceptance-based perspective on the update semantics. Second, the truth-based notion of consequence is strictly stronger than the acceptance-based notion, i.e., if Γ entails φ in the truth-based notion, Γ also entails φ in the acceptance-based notion, but not vice versa. The truth-based notion of consequence thus provides a more expressive perspective on the update semantics; again, the update semantics can plausibly be coupled with a variety of different logics, but I think for present

purposes it makes sense to couple it with a logic which does not wash out any of its expressive power, and so I will stick with the truth-based notion.[20] Readers interested in the perspective gained from an acceptance-based vantage point on the update semantics are, again, referred to the discussion of the state-based semantics below.

With this background, we can gloss the usual update semantic rules for atomic sentences and $\Diamond$ as follows. For any model $\mathfrak{u} \in \mathcal{U}$, an atomic sentence p is $\text{true}_\mathfrak{u}$ at $\langle s, c\rangle$ just in case c is the result of removing all worlds w from s such that $v_\mathfrak{u}(p, w) = 0$, where $v_\mathfrak{u}$ is $\mathfrak{u}$'s valuation function. And a sentence of the form $\Diamond\varphi$ "tests" whether φ is compatible with the input context: $\Diamond\varphi$ is $\text{true}_\mathfrak{u}$ at $\langle s, c\rangle$ just in case, roughly, s includes a φ-world and $s = c$, or s doesn't include a φ-world, and $c = \varnothing$; more precisely, just in case $[\![\varphi]\!]_\mathfrak{u}(s) \neq \varnothing$ and $s = c$, or else $[\![\varphi]\!]_\mathfrak{u}(s) = \varnothing$ and $c = \varnothing$.

We first show that the update semantics is no more expressive than the relational semantics with respect to $\mathcal{L}^\Diamond$:

Fact 3.6. $\mathcal{U} \preceq_{\mathcal{L}^\Diamond} \mathcal{R}$.

The proof goes by way of constructing an injection from pairs of contexts to relational indices which preserves truth for all sentences in $\mathcal{L}^\Diamond$. The injection goes by way of distinguishing a variety of different cases, and is not particularly intuitive, which makes Fact 3.6 somewhat surprising. We can also show that the relational semantics is more expressive than the update semantics:

Fact 3.7. $\mathcal{R} \not\preceq_{\mathcal{L}^\Diamond} \mathcal{U}$.

The proof is identical to the proof of Fact 3.3 (that $\mathcal{R} \not\preceq_{\mathcal{L}^\Diamond} \mathcal{D}$). Thus we have:

Fact 3.8. $\mathcal{U} \prec_{\mathcal{L}^\Diamond} \mathcal{R}$.

Interestingly, though, we do not have parallel results for the domain semantics:

Fact 3.9. $\mathcal{U} \not\preceq_{\mathcal{L}^\Diamond} \mathcal{D}$.

Fact 3.10. $\mathcal{D} \not\preceq_{\mathcal{L}^\Diamond} \mathcal{U}$.

We can prove these facts by showing that there is no truth-preserving injection from the indices of $\mathcal{U}$ to those of $\mathcal{D}$, and *vice versa*.[21] Thus the domain and update semantics are expressively incommensurable—though both are strictly less expressive than the relational semantics.

[20] It is also worth noting that not everything in the update framework is definable in terms of acceptance—in particular, the update treatment of 'might' cannot be reformulated in an equivalent way just in terms of the fixed points of its complement. To see this, let $Acc(\varphi)$ be the set of φ's accept states, i.e. $Acc(\varphi) = \{c : [\![\varphi]\!]_\mathcal{U}(c) = c\}$. In standard extensions of the update semantics to conjunction and negation, for atomic p, $Acc(p \wedge \neg p) = Acc(\Diamond p \wedge \neg p) = \{\varnothing\}$: these sentences have all the same fixed points, namely, just $\varnothing$. But they interact differently with $\Diamond$: e.g. $\Diamond(p \wedge \neg p)$ is only accepted by $\varnothing$, whereas $\Diamond(\Diamond p \wedge \neg p)$ is accepted by some non-empty sets, namely those which include both p- and $\bar{p}$-worlds; and so the update semantics for $\Diamond$ cannot be defined just in terms of the accept states of its complement (though see Gillies 2018 for a variant which can be).

[21] These facts are particularly interesting in relation to the limited equivalence between the update and domain semantics proved in Rothschild 2017. See also Rothschild and Yalcin 2015, 2016 for detailed discussion of dynamic semantics and their relation to static semantics.

Note, finally, that these relations are all preserved for $\mathcal{L}^{\Diamond-}$, the language which does not allow stacked 'might's. First, we have $\mathcal{D} \not\preceq_{\mathcal{L}^{\Diamond-}} \mathcal{U}$ (the proof is the same as the proof of Fact 3.10). That, together with Fact 3.5 (that $\mathcal{D} \prec_{\mathcal{L}^{\Diamond-}} \mathcal{R}$), and the fact that $\preceq$ is transitive (see Fact 3.13 below), shows that $\mathcal{R} \not\preceq_{\mathcal{L}^{\Diamond-}} \mathcal{U}$; otherwise, we would have $\mathcal{D} \preceq_{\mathcal{L}^{\Diamond-}} \mathcal{U}$. Finally, the proof that $\mathcal{U} \preceq_{\mathcal{L}^{\Diamond}} \mathcal{R}$ extends immediately to $\mathcal{L}^{\Diamond-}$, so we have:

Fact 3.11. $\mathcal{U} \prec_{\mathcal{L}^{\Diamond-}} \mathcal{R}$.

The last system I will discuss here is the *state-based* semantics of Hawke and Steinert-Threlkeld 2016.[22] The state-based semantics $\mathcal{S}$ is very similar to the update semantics, except an index is a single set of worlds. At any model $\mathfrak{s} \in \mathcal{S}$, an atomic sentence p is true at a set of worlds c just in case, for every world $w \in c : v_{\mathfrak{s}}(p, w) = 1$. And a sentence of the form $\Diamond\varphi$ is true in $\mathfrak{s}$ just in case φ is true in $\mathfrak{s}$ in some $\{w\} \subseteq c$. The state-based semantics is strictly less expressive than the domain semantics:

Fact 3.12. $\mathcal{S} \prec_{\mathcal{L}^{\Diamond}} \mathcal{D}$.

Furthermore, for any language $\mathcal{L}$, $\preceq_{\mathcal{L}}$ is transitive across semantic theories isomorphic over $\mathcal{L}$; this follows from Fact 3.13:

Fact 3.13. For any language $\mathcal{L}$ and class of semantic theories $\mathfrak{T}$ isomorphic with respect to $\mathcal{L}$, $\preceq_{\mathcal{L}}$ is a partial pre-order over $\mathfrak{T}$.

So we also have:

Fact 3.14. $\mathcal{S} \prec_{\mathcal{L}^{\Diamond}} \mathcal{R}$.

3.3 Quantified modal languages

The final comparison of expressive power which I will explore here concerns quantified modal languages. This is a particularly interesting topic for present purposes because epistemic modals embed in fascinating ways in the scope of quantifiers—ways that have been used to argue against the standard semantics and in favor of various quantified enrichments of the update semantics. The present expressibility results can, however, be extended to show that, if we extend all these frameworks to quantificational languages in the most straightforward way, the expressive hierarchies for the non-quantified language remain unchanged.

To show this, consider a standard quantificational language $\mathcal{L}_v^{\Diamond}$, built out of a vocabulary comprising variables $x_i : i \in \mathbb{N}$; n-place relation symbols $R_i^n : i \in \mathbb{I}$ for every $n \geq 0$; and a one-place sentence operator $\Diamond$. $\mathcal{L}_v^{\Diamond}$ is the smallest set comprising atomic sentences with the form of an n-place relation

[22] It is a more complicated matter to extend the comparison to the bilateral state-based semantics in Aloni 2016, Steinert-Threlkeld 2017, Chapter 3.

symbol R^n followed by an n-tuple of variables; and sentences of the form $\Diamond\varphi$, for any $\varphi \in \mathcal{L}_v^{\Diamond}$. We can extend the basic relational semantic theory $\mathcal{R}$ to a semantic theory $\mathcal{R}^{\exists}$ for this quantified language by adding a domain to the semantic theory, and adding a variable assignment (a function from variables to elements of the domain) to our indices, so that they amount to triples comprising a variable assignment, accessibility relation, and world. Our valuation function now takes a world and an n-place relation symbol and returns an n-place relation on the domain. Our truth clauses for atomic sentences and $\Diamond$ will be generalized in the usual way, and our semantics for $\Diamond$ remains essentially unchanged.[23] We can treat quantifiers as sentence operators (shifting variable assignments, in the usual way) which we can freely add to our language. We can likewise enrich the domain semantics to a semantic theory $\mathcal{D}^{\exists}$ of $\mathcal{L}_v^{\Diamond}$ in a parallel fashion, augmenting our indices so they are triples of variable assignments, information states, and worlds, and extending the interpretation function in the obvious way. It is clear that these changes do not affect the expressive hierarchy between the domain and relational semantics. That is, we have:

Fact 3.15. $\mathcal{D}^{\exists} \prec_{\mathcal{L}_v^{\Diamond}} \mathcal{R}^{\exists}$.

The proof is a straightforward generalization of the proof of the parallel result for the non-quantified case.

Things get more interesting when we turn to the update semantics.[24] The most obvious way to incorporate quantification into the update semantics is to simply treat intensions as functions from a variable assignment to a function from contexts to contexts (as in Yalcin 2015), so that $\mathcal{U}^{\exists}$-indices have the form $\langle a, \langle s, c\rangle\rangle$, for a a variable assignment and s and c contexts.[25] For any model $\mathfrak{u}^{\exists} \in \mathcal{U}^{\exists}$, we say that, for atomic sentence p of the form $R^n(\langle x_1, x_2, \ldots x_n\rangle)$, p is $\text{true}_{\mathfrak{u}^{\exists}}$ at $\langle a, \langle s, c\rangle\rangle$ iff c is the result of removing from s all and only worlds w such that $\langle a(x_1), a(x_2), \ldots a(x_n)\rangle \notin v_{\mathfrak{u}^{\exists}}(R^n, w)$. And we say that $\Diamond\varphi$ is $\text{true}_{\mathfrak{u}^{\exists}}$ at $\langle a, \langle s, c\rangle\rangle$ iff $[\![\varphi]\!]_{\mathfrak{u}^{\exists}}(a)(s) \neq \varnothing \wedge s = c$, or $[\![\varphi]\!]_{\mathcal{R}^{\exists}}(a)(s) = \varnothing = c$. If we go this way, then, once again, the expressive hierarchy from above will be preserved: the quantified update semantics will be strictly less expressive than the quantified relational semantics.

Fact 3.16. $\mathcal{U}^{\exists} \prec_{\mathcal{L}_v^{\Diamond}} \mathcal{R}^{\exists}$.

[23] I.e. for any model $\mathfrak{r}^{\exists} \in \mathcal{R}^{\exists}$, where $\langle a, f, w\rangle$ is a relational index, with a a variable assignment, f a modal base, and w a possible world, we say that an atomic sentence of the form $R^n(\langle x_1, x_2, \ldots x_n\rangle)$ is $\text{true}_{\mathfrak{r}^{\exists}}$ at $\langle a, f, w\rangle$ iff $\langle a(x_1), a(x_2), \ldots a(x_n)\rangle \in v_{\mathfrak{r}^{\exists}}(R^n, w)$, and $\Diamond\varphi$ is $\text{true}_{\mathfrak{r}^{\exists}}$ at $\langle a, f, w\rangle$ iff there is a world $w' \in f(w)$ such that φ is $\text{true}_{\mathfrak{r}^{\exists}}$ at $\langle a, f, w'\rangle$.

[24] I'll set aside the state-based semantics for now; I don't know of quantificational versions of that semantics.

[25] Interestingly, this is not the most common way to incorporate quantification into the dynamic framework in which the update semantics is cast. In the standard quantified extension of that framework, developed in Heim 1982, 1983, indices are treated, not as pairs of a variable assignment and a pair of information states, but rather as pairs of 'files', where a file is a set of world/variable assignment pairs. I will not explore the expressive hierarchies that would result from this more complicated way of going.

4 The upshot

There is much more to explore here, but I will stop for the present. Let me briefly summarize the situation, and then emphasize a few upshots of this discussion. The (quantified) domain semantics, (quantified) update semantics, and state-based semantics are all strictly less expressive than the (quantified) relational semantics with respect to a language comprising 'might' and atomic sentences. In other words, for any way of extending this language with a new set of sentence operators and any way of extending the domain, update, or state-based semantic theories to give semantics for those operators, we can extend the relational framework to these operators in a way which exactly replicates the logic of that operator in the given semantics. But the converse is not true: there are operators definable in the relational framework whose logic cannot be replicated in the domain, update, or state-based frameworks. This shows that there are no data involving 'might' embedded under some operator which the domain, update, or state-based semantics can make sense of, but which the relational semantics cannot make sense of.[26]

This, in turn, sheds new light on the debate about the semantics of epistemic modals. At a high level, that dialectic has sometimes been presented roughly as follows. First, a conservative assumption in favor of the relational semantics is generally taken for granted. Then, this assumption is challenged with data involving 'might' embedded under other operators which the relational semantics putatively has trouble accounting for. Finally, the data are used to advocate a revisionary semantics which can make sense of them when extended in an appropriate way to give a semantics for the embedding operator.

The present results show that this gets things backwards. On the one hand, these results show that there are no embedding data which on their own tell against the relational semantics and in favor of the domain, update, or state-based semantics. That is, for any embedding data which we can account for within the domain, update, or state-based semantics by extending those semantics to the embedding operator(s) in question, we can account for the same data in the relational semantics by extending the relational semantics in a way which matches the logic of the new operator(s) in the

[26]Of course, this holds only as long as we focus on sentences where 'might' takes as a complement just sentences in our simple starting language. In fact, of course, 'might' can embed sentences of much more complexity. But, as far as the embedding behavior of epistemic modals goes, this limitation seems to be harmless; all of the data that I know of which have been used to motivate departures from the relational semantics stay within these bounds. This limitation follows from the assumption that $\mathcal{L}^{\mathcal{O}}$ is closed only under the operators in $\mathcal{O}$, and not under the operators in the vocabulary of $\mathcal{L}$. If we made the latter assumption, we would avoid this limitation, but the proofs below would not go through. Note that many (but not all) of the results above can be easily extended to more complex starting languages. (See brief discussion in Footnote 18. In some cases, no such extension will be possible, e.g. if we compare the update and relational semantics with a starting language which includes conjunction, given its standard entries in these two frameworks. But the comparison of that result to those above shows that the resulting failure has to do with those entries for conjunction, not with the semantics for epistemic modals; the results above show that the logical features of the standard update conjunction can be replicated in the relational framework.)

domain/update/state-based frameworks. But the converse does not hold, and so there may well be data which tell the other way: embedding data which the relational framework can make sense of, but which the domain, update, and state-based frameworks cannot make sense of.

On the other hand, excessive expressive power is to be avoided in giving a semantics for natural language. So these results also provide the basis for an argument in favor of the revisionary, less expressive alternatives. If natural language does not make use of the full expressive power of the relational semantics, we may want to record this fact in our semantic theory. The less expressive a theory of 'might' is, the fewer stipulations are needed to make it match data from natural language. Thus, if the less expressive frameworks can make sense of the behavior of embedded epistemic modals, then the present expressibility results show that there should be a simplicity-based prejudice in their favor. This makes precise the intuition reported by Ninan and others that there is something about the domain semantics which makes it simpler than the relational semantics.

These considerations are, of course, only *pro tanto*. There are many other considerations that can bear on the choice between semantic theories. For instance, one framework may be able to account for embedding data in a more natural way than another, (for instance, by doing so in a computationally simpler way; it is important to note that just because $\mathcal{T} \preceq_{\mathcal{L}} \mathcal{T}'$, it does not follow that $\mathcal{T}'$ can match the logic of any extension of $\mathcal{T}$ *in as computationally simple way as* $\mathcal{T}$ *does*).[27]

Let me put all this a bit more generally, to bring out the utility (and limitations) of comparisons of potential expressibility. Consider a debate between two semantic theorists about how to make sense of some fragment of natural language. The first theorist may point to the logical behavior of sentences in some extension of this fragment as evidence in favor of her theory. The second theorist could respond by showing that there exists an extension of her preferred theory to that extension which also captures the logical behavior in question. But the present considerations show that a more general response is also available to her. If she can show that her theory has greater potential expressibility than the first theorist's, then she can show that the first theorist can't *ever* win an argument so easily: whatever extension the first theorist offers up, the second theorist will be able to match the logic of that extension. This does not decide which theory is correct. For one thing, this very fact, although it undermines the first theorist's original argument for her theory, may be taken as a different point in favor of the first theory, as it shows that there is a precise sense in which the first theory is less flexible than the second. Second, there may be other, independent considerations which bear on the choice between the theories. Thus considerations of relative potential expressibility are not decisive, but they play at least two important roles: they help us determine when an argument based purely on the logic of some extension of a semantic theory can *on its own* be an argument in favor of that theory

[27] Cf. debates in syntax, where more powerful grammars are sometimes preferred simply because of the ease with which they can capture the data; thanks to Roni Katzir for pointing out this connection.

and against another; and they help us make precise intuitions about the relative simplicity of theories.

5 The domain versus relational semantics

This concludes my abstract discussion of relative potential expressibility. My plan for this final part of the paper is to briefly descend from these abstract considerations back into the more concrete debate about which theory is best able to account for epistemic modals in natural language. My goal is not to choose between these theories here. I will instead focus on a limited subset of embedding data, making a case study of the behavior of epistemic modals under attitude predicates. The goal will be to bring out the kinds of methodological considerations that should be brought to bear in deciding between different semantic theories in light of expressibility results like those proved above.[28]

As we saw above, Ninan (2016) shows that we can account for Yalcin's data in the relational system. More generally, Ninan's semantics for attitude predicates, plus the relational semantics, is in fact equivalent to Yalcin's system.[29] In light of this equivalence, together with the expressibility results which showed that there is a sense in which the domain semantics is more constrained than the relational semantics, it would be tempting to conclude that the domain semantics is to be preferred over the relational semantics, at least *modulo* considerations about epistemic modals in other environments. But this would be too fast, because it turns out the predictions of Yalcin's and Ninan's systems are problematic.[30]

These systems have at least three serious empirical problems. First, as Dorr and Hawthorne (2013) point out, Yalcin's system (and hence Ninan's system) predicts that, for non-modal φ and agent A, the inference from φ to $\ulcorner$A knows might $\varphi\urcorner$ will be valid: on these accounts, the latter is true just in case the truth of φ is compatible with A's knowledge, and any truth is compatible with anyone's knowledge. But this is obviously wrong. It is possible to fail to know that some truth might obtain; this is just what happens when one has a false belief. Suppose John sees Mark enter his office and close the door. Unbeknownst to John, Mark has a secret exit in the floor of his office, and has used this exit to leave the office and go to the bar. In this situation, 'John knows that Mark might be at the bar' seems plainly false. But, since Mark is in fact at the bar, the Yalcin/Ninan approach wrongly predicts this to

[28] In Mandelkern 2019, I explore a much wider variety of embedding data, and give a different response to those data. To be clear, I stand by the response I give there; the proposals I explore here are local and *ad hoc* in an entirely unsatisfying way, and serve a merely illustrative purpose. The 'bounded theory' that I develop there builds on the relational theory. However, as I point out there (Footnote 58), a similar response could be given in a domain framework. For what it's worth, my own view is that the relational framework is preferable, but for reasons having to do with higher-level considerations about communication (which I discuss in Mandelkern (2018a)), not on the basis of embedding data.

[29] I.e. for any attitude verb V, world w, information state s, and modal base f, $\ulcorner$A V's $\varphi\urcorner$ is true in Yalcin's system at an index $\langle s, w\rangle$ just in case it is true in Ninan's system at the index $\langle f, w\rangle$.

[30] The same problems face the update semantics, when combined with the approach to attitudes in Heim 1992, or the event-relative modal and attitude semantics of Hacquard 2006, 2010.

be true.[31]

The second problem is that the Yalcin/Ninan framework gets the entailment relations between ⌜A knows might φ⌝ and ⌜A believes might φ⌝ backwards. In this framework, the first says that the truth of φ is compatible with A's knowledge; the latter that it is compatible with A's beliefs (provided A's beliefs are consistent). Since whatever is compatible with someone's beliefs is compatible with their knowledge, but not *vice versa*, this framework predicts that ⌜A knows might φ⌝ does *not* entail ⌜A believes might φ⌝; but that, as Dorr and Hawthorne (2013) point out, that ⌜A consistently believes might φ⌝ *does* entail ⌜A knows might φ⌝. But this is wrong: just as in the non-modal case, likewise in the modal case, knowledge entails belief, but not *vice versa*. Birthers consistently believe that Obama may have been born in Kenya, but they don't know this. By contrast, if John knows Mark may be in his office, then he also believes this; 'John knows Mark might be in his office, but he doesn't believe Mark might be in his office' has the same sense of incoherence as in the non-modal case.[32]

The final problem for the Yalcin/Ninan system is that it predicts that 'must' is vacuous under attitudes.[33] That is, sentences with the form ⌜A V's must φ⌝ and ⌜A V's φ⌝ are predicted to be semantically equivalent: modals embedded under an attitude verb are predicted only to contribute quantificational force, and since the universal quantificational force of 'must' matches the universal quantificational force of the embedding predicate, 'must' is predicted to have no effect. This prediction, however, is wrong. A construction with the form ⌜A knows/believes must φ⌝ is generally only felicitous if A's evidence for the truth of φ is in some sense indirect.[34] For instance—to modify a stock example—suppose that Sue is watching it rain, and on this basis concludes that it's raining out. Then we can say 'Sue knows/believes it's raining', but the 'must'-variant 'Sue knows/believes it must be raining' is quite odd. By contrast, suppose that Sue can't see outside, but sees some of her colleagues come inside with wet umbrellas, and on this basis concludes that it's raining out. Then either the 'must' or non-modal variant is acceptable. There are many kinds of explanation we might seek out for

[31] Yalcin (2012) discusses this problem in the context of the update framework. A referee for this journal helpfully points out that this prediction of the Yalcin/Ninan system would be palatable if there were *some* reading of 'knows' on which the inference from φ to ⌜A knows might φ⌝ looks valid. If there were, then sequences like the following should have a coherent reading: 'Susie is completely convinced that it's sunny out; she is, after all, looking out at what appears to be a sunny sky. But she knows that it might be raining out, because in fact it's raining out, and the apparently sunny sky is just a clever projection.' On any reading, the last sentence here sounds like a non sequitur.

[32] See Hawthorne et al. (2016); Bledin and Lando (2017); Beddor and Goldstein (2018) for discussion of related cases. A referee for this journal helpfully points out that these first two problem stem from the same logical feature of the Yalcin/Ninan system: that on that system, ⌜A consistently believes/knows φ⌝ is equivalent to ⌜φ is consistent with A's beliefs/knowledge⌝. These problems are, nonetheless, distinct, and are important to keep separate because it is possible to solve one problem without solving the other (as we will see presently).

[33] Assuming 'must' is defined as the dual of 'might', with negation given its standard Boolean meaning. This follows from the more general problem that to accept φ and to accept ⌜Must φ⌝ amount to the same thing in this system, a problem Hacquard 2010, §6.1.2 discusses.

[34] This generalizes a common parallel observation about unembedded 'must'; see Karttunen 1972; Veltman 1985; Kratzer 1991; von Fintel and Gillies 2010; Kratzer 2012; Matthewson 2015; Lassiter 2016; Giannakidou and Mari 2016; Sherman 2018; Mandelkern 2017b, 2018b. For specific discussion of embedded 'must' see Ippolito 2017.

this difference, including broadly pragmatic explanations, but it is hard to see how the Yalcin/Ninan systems could provide the foundation for any such explanation, since, on this approach, the 'must' and non-modal variants are, again, semantically equivalent.[35]

The Yalcin/Ninan system is thus not empirically viable. Yalcin's and Ninan's system show us how to respond to Yalcin's data within a domain and relational semantics, respectively. In both cases, those responses have implausible results. The expressibility results above showed that the relational semantics has more expressive power than any of the alternative views we explored. The implausibility of the Yalcin/Ninan system may appear to show that we need an even more expressive semantics than the relational one in order to have the flexibility to account for Yalcin's data in a plausible way.

But this would be too fast. The main point which I wish to make in this section is that reasoning like this would only be valid if Yalcin's and Ninan's systems were the *weakest* ways one could possibly account for Yalcin's data in the domain and relational frameworks, respectively. If this *were* so, then things would look very bad indeed for those frameworks: we would know that any way those frameworks could account for Yalcin's data will validate the implausible inferences just reviewed. This would, in turn, provide sufficient motivation to reject those theories and pursue a new, more expressive theory of epistemic modals. But, by contrast, if Yalcin's and Ninan's systems do *not* represent the weakest ways to respond to Yalcin's data within the domain and relational frameworks, respectively, then the failure of the Yalcin/Ninan systems tells us nothing about the viability of those frameworks. More generally, my point is the following: when faced with new data, the only way to show that a semantic system is incapable of making sense of those data is by looking at the *weakest* way that system can account for the data. It is only if the weakest possible account of the data within that system validates implausible entailments that we know we need a more expressive underlying system, rather than simply a different way of responding to the data within that system.

And it turns out that the Yalcin/Ninan system is not the weakest way to account for Yalcin's data, within either the relational or domain frameworks. In both cases, there are much weaker constraints available, which account for Yalcin's data and avoid the problems just enumerated. To show this,

[35] It may look as though these data can be explained in a relatively conservative way within a close variant of the Yalcin/Ninan approach by adopting a presupposition of indirectness along the lines suggested by von Fintel and Gillies (2010). But, first, to do this, we would have to depart substantively from the domain semantics, since information states do not provide enough structure to formulate a presupposition like the one von Fintel and Gillies propose; on that proposal, modals are evaluated relative to a set of propositions (representing an agent's direct evidence), and presuppose that no single element of the set entails the modal's prejacent or its negation. Even if we modify the domain semantics so that the von Fintel and Gillies proposal is statable, Ippolito (2017) gives convincing arguments that indirectness does not project like a presupposition, and so should not be encoded as a presupposition at all (nor will we have better luck encoding it as a conventional implicature, which would face the same objections). I am inclined to think instead that a pragmatic view like the one that I develop in Mandelkern (2017b, 2018b) is more plausible (cf. Degen et al. 2015). But that approach crucially requires that there be a difference in truth conditions between 'must' sentences and corresponding non-modal sentences, and so cannot get off the ground if ⌜A V's must φ⌝ is semantically equivalent to ⌜A V's φ⌝. There are, of course, pragmatic accounts which distinguish sentences which are semantically equivalent, but I do not see a natural account to apply to the present case. In any case, the first two two points alone provide motivation to look for alternatives.

I will explore what the weakest constraint is in each framework which accounts for these data. To make this discussion tractable, I will make two background assumptions: first, that our connectives are the standard Boolean ones;[36] second, that attitudes are represented as sets of possible worlds, in the broadly Hintikkan framework that Ninan and Yalcin both assume—i.e. that attitude predicates have as their core semantic values universal quantifiers over accessible worlds. There may be reasons to relax both these assumptions, but they facilitate the present discussion, and are harmless as far as present purposes are concerned.

Given all this, the weakest constraint in the relational framework which suffices to account for Yalcin's data—that is, to ensure that ⌜A supposes (φ and might not φ)⌝ entails that A's suppositions are inconsistent—is the following: provided A's suppositions are consistent, then some supposition world can only access other supposition worlds. We can schematically encode this constraint, which I'll call the *subset relational constraint*, as follows:

Definition 5.1. *Subset attitude semantics, relational version*:

- $[\![\text{A V's } \varphi]\!]^{f,w}$

  a. defined only if $V_{A,w} \neq \varnothing \rightarrow \exists w' \in V_{A,w} : f(w') \subseteq V_{A,w}$ *subset relational constraint*

  b. where defined, true iff $\forall w' \in V_{A,w} : [\![\varphi]\!]^{f,w'} = 1$ *standard Hintikkan truth conditions*

To see that this constraint is necessary and sufficient to account for Yalcin's data in the relational framework, given our background assumptions, first assume that ⌜A supposes (φ and might not φ)⌝ is true; then we know that (i) all of A's suppositions worlds make φ true (relative to f), and (ii) all of A's suppositions worlds can access a world where φ is false (relative to f). But the subset relational constraint ensures that, if A's suppositions are consistent, then one of A's supposition worlds can only access other supposition worlds. It follows that (i) and (ii) are only both satisfied when there are no worlds compatible with A's suppositions. Note next that the subset relational constraint is *necessary* to account for Yalcin's data, given our background assumptions: if the constraint is violated, then ⌜A supposes (φ and might not φ)⌝ will not entail that A's suppositions are inconsistent. Suppose that A's suppositions are consistent at w, and the subset relational constraint is violated for some modal base f: that is, $S_{A,w}$ is non-empty, and $\forall w' \in S_{A,w} : f(w') \setminus S_{A,w} \neq \varnothing$. Let φ denote $S_{A,w}$.[37] By construction, every world in $S_{A,w}$ is a φ-world, and every world in $S_{A,w}$ will be able to access a world under f where φ is false. Thus ⌜A supposes (φ and might not φ)⌝ is true at w, even though A's suppositions at w are consistent.

[36] While adopting non-standard connectives could on its own explain Yalcin's data (as Rothschild and Klinedinst (2015); Mandelkern (2019) discuss) it would not on its own account for nearby order and scope variants.

[37] It will generally be possible to come up with a sentence that denotes the content of an attitude state in natural language: e.g. just let $\varphi =$'What A V's at w'.

Thus, given our background assumptions, the subset relational constraint is the weakest constraint which accounts for Yalcin's data in the relational framework. The first thing to note about this constraint is that it is much weaker than the constraint implicit in Ninan's system—that modal bases must be constant functions to the set of attitude world. And thus the relational semantics is by no means locked into the Yalcin/Ninan approach if it is to make sense of Yalcin's data.

Before exploring the plausibility of the subset relational system (the system that results from the relational semantics together with the subset attitude semantics), let us explore the parallel question for the domain semantics. It turns out that there is no way to exactly replicate the subset relational system in the domain framework, given our background assumptions:[38] something which is unsurprising, given the greater expressive power of the relational framework, and which provides a helpful concrete illustration of those expressibility results. But neither is the domain semantics locked into the Yalcin/Ninan framework. The weakest constraint in the domain framework which accounts for Yalcin's data is the following: a modal in the complement of ⌜A supposes⌝ is always evaluated relative to a *subset* of A's supposition worlds. A simple way to encode this constraint, which I call the *subset domain constraint*, is as follows:[39]

Definition 5.2. *Subset attitude semantics, domain version:*

- $[\![\text{A V's } \varphi]\!]^{s,w}$
  - a. defined only if $s \subseteq V_{A,w}$ *subset domain constraint*
  - b. where defined, true iff $\forall w' \in V_{A,w} : [\![\varphi]\!]^{s,w'} = 1$ *standard Hintikkan truth conditions*

Given this constraint, if ⌜A supposes (φ and might not φ)⌝ is true at $\langle s, w\rangle$, then (i) all the worlds compatible with A's suppositions in w make φ true (relative to s); and (ii) all the worlds compatible with A's suppositions are such that some world in s makes φ false (relative to s). These two conditions can only be jointly met if there are no worlds compatible with A's suppositions. Next suppose that A's suppositions are consistent at w, and that the subset domain constraint is violated for some information state s, i.e. $s \setminus S_{A,w} \neq \varnothing$. Let φ denote $S_{A,w}$. Then ⌜A supposes (φ and might not φ)⌝ will be true at $\langle s, w\rangle$, since (i) all of A's supposition world are φ-worlds; and (ii) some world in s makes φ-false, since $s \setminus S_{A,w}$ is non-empty and by construction includes only ⌜$\neg\varphi$⌝-worlds. So, if the subset domain constraint is violated, ⌜A supposes (φ and might not φ)⌝ can be true even if A's suppositions are consistent.

[38] The proof follows from the fact that, given our background assumptions and given any semantics for attitudes, the domain framework will validate the inference from ⌜A V's might (φ or ψ)⌝ to ⌜(A V's might φ) or (A V's might ψ)⌝; but this inference is invalid in the subset relational framework. See Stalnaker 1984; Yalcin 2011 for criticism of one prediction of this inference pattern: that ⌜A believes might φ or A believes might $\neg\varphi$⌝ is a logical truth.

[39] A different route to an approach which ends up being essentially equivalent to the subset domain semantics goes by adding an ordering source to Hacquard (2010)'s event-relative semantics; see Hacquard 2010, §6.1.2 for a brief discussion of this possibility.

Thus, given our background assumptions, the subset domain constraint is the weakest constraint which accounts for Yalcin's data in the domain framework. And, once more, this constraint is much weaker than the one relied on in Yalcin's framework—that modals under attitudes are evaluated relative to exactly the set of attitude worlds. Thus the domain semantics, like the relational semantics, is not locked into the Yalcin/Ninan approach if it is to make sense of Yalcin's data.

The key question at this stage is whether the subset relational or subset domain systems improves over the Yalcin/Ninan systems. Strikingly, both do (albeit to varying degrees). First, both approaches predict that non-modal φ does *not* entail ⌜A knows might φ⌝, avoiding the most serious problem for the Yalcin/Ninan system: this is because, on both approaches, ⌜A knows might φ⌝ is strictly stronger than ⌜φ is compatible with A's knowledge⌝. Second, both approaches, unlike the Yalcin/Ninan approach, predict that ⌜A knows might φ⌝ Strawson entails ⌜A believes might φ⌝ (in the terminology of von Fintel 1997): whenever both sentences are well-defined, if the first is true, the second is as well (holding fixed the modal base/information state parameter). The subset relational approach also correctly predicts that ⌜A consistently believes might φ⌝ does not entail ⌜A knows might φ⌝; by contrast, the subset domain approach still wrongly predicts this entailment is Strawson valid. Finally, both approaches predict that ⌜A V's must φ⌝ and ⌜A V's φ⌝ are not semantically equivalent: the subset relational approach predicts that neither entails the other, while the subset domain approach predicts that the latter entails the former, but not *vice versa* (in both cases, there is much more to be done to marshal these facts into an explanation of how 'must' patterns under attitude verbs, but this is a clear improvement over the Yalcin/Ninan prediction that ⌜A V's must φ⌝ is semantically equivalent to ⌜A V's φ⌝).

The subset domain system is less attractive than the subset relational system with respect to its predictions about the relation between belief and knowledge. But there is much more to explore before we come to any conclusion here; in particular, we must, among other things, explore alternative approaches which relax the two background assumptions we have made here.[40] The goal of the present excursus is not to decide between the domain and relational approaches, but rather to emphasize that, in deciding whether some embedding data show that a given semantic framework is not expressively powerful enough to make sense of natural language, we must always explore the *weakest* constraint within that framework which makes sense of those data. It is only if that constraint makes implausible commitments that the data truly tell against the framework in question. In the present case, these considerations showed that there are much weaker ways to account for Yalcin's data in both the domain and relational framework than with the implausible Yalcin/Ninan framework. These weaker

[40] Again, in Mandelkern (2019), I explore a system which abandons Boolean semantics for the connectives. In a different direction, Beaver 1992, 2001; Rothschild 2011; Yalcin 2012; Willer 2013 develop systems which avoid the first two of these problems by treating attitude states, not as sets of worlds, but as sets of sets of worlds. Those approaches avoid the first and second problems raised here, though they still face the third problem: they predict that ⌜A V's must φ⌝ is semantically equivalent to ⌜A V's φ⌝, at least within any eliminative fragment like the ones they are working with.

approaches are both much more plausible than the Yalcin/Ninan system, and thus these embedding data do not, after all, tell against either framework as a candidate semantics for epistemic modals.

Let me conclude by pointing to a more indirect upshot of this discussion. The domain semantics has a substantially stronger logic than the relational semantics; in particular, enriched with standard classical treatments of the connectives, the domain semantics validates the logic **K45** (or **S5**, if we also impose a reflexivity constraint by assuming that, for all domain indices $\langle s, w\rangle$, $w \in s$)[41] whereas the relational system validates just the much weaker logic **K** (again, unless we assume a reflexivity constraint (as I have not), in which case we would have a **T** logic). Interestingly, Ninan's system, although it adopts the relational semantics, still validates **K45** in a local sense: the axioms of **K45** will be valid in the scope of an attitude verb. This may make it look as though Yalcin's data are really an argument for **K45**: we could validate **K45** directly by adopting the domain semantics, or indirectly (in a local way) by adopting Ninan's semantics, but either way, we must validate **K45**, at least in the scope of attitude predicates, if we are to make sense of the data. But the subset relational approach shows this is wrong: it is mere happenstance that the most prominent treatments of the data both validate this strong logic (at least in the scope of attitude predicates), for the subset relational approach does not validate **K45** even for modals in the scope of attitude verbs (e.g. on this approach, ⌜A believes might p⌝ can be true without ⌜A believes must (might p)⌝ being true, and ⌜A believes must p⌝ can be true without ⌜A believes must (must p)⌝ being true). And so we can make sense of Yalcin's data without validating **K45** even in a local sense. This shows that—even though (I have argued) there is an argument for the domain semantics on the basis of its *expressive weakeness*—there is no argument from Yalcin's data for the domain semantics on the basis of *its stronger logic*, which turns out to be strictly orthogonal to accounting for those data. (There may, of course, be other arguments that we want a stronger logic for epistemic modals than **K**; my present point is simply that the present discussion shows that Yalcin's embedding data do not provide such arguments.)

6 Conclusion

I have argued that we can gain new insight into controversies about the semantics of natural language expressions by taking an abstract perspective on the potential expressive power of different semantic theories. I have developed a formal framework to make these comparisons precise, and have used this framework to explore the relative potential expressive power of different semantics for epistemic modals. These comparisons show that, for any embedding operator which can be defined in the domain, update, or state-based semantics, a corresponding operator can be defined in a relational framework, but not conversely. This shows that the dialectic in this debate is roughly the opposite of

[41] See Schultz 2010; Holliday and Icard 2017.

what it has often been taken to be. On the one hand, the relational theory can do anything that these revisionary theories can do, showing that it is not defenders of the relational framework who must fight rearguard actions when new data are discovered, but rather defenders of the revisionary frameworks who must show that those frameworks have the expressive power to account for those data. But, on the other hand, the relative expressive weakness of the domain, update, and state-based semantics provides powerful *pro tanto* considerations in their favor. In the last part of the paper, I explored the comparison between the relational and domain frameworks in light of these results, focusing on the behavior of modals under attitudes. I emphasized there the importance of methodological parsimony in choosing between semantic frameworks: the only way to see whether a given semantic framework has the expressive power to account for some new domain of data is by finding the *weakest* way the framework can do so, and then asking if the result makes implausible commitments.

I have focused on epistemic modals here because they provide an apt illustration of the utility of the expressive comparisons I have developed. While I hope this discussion has advanced our understanding of the meaning of epistemic modals, my broader goal has been to develop a formal framework with widespread application in semantics. Semantic theory has often advanced thanks to results regarding expressive power—for instance in the theory of tense and temporal adverbs,[42] or of generalized quantifiers.[43] The framework for comparing relative potential expressibility introduced here makes precise a new kind of question we can ask about the expressive power of different semantic frameworks, and the characterization result proved shows how to straightforwardly answer those questions. There is much more work to do in exploring the applications of this framework, as well as exploring and extending the underlying formalism. Of particular interest is a set of questions about computational complexity: if $\mathcal{T} \preceq_{\mathcal{L}} \mathcal{T}'$, that means that, for any operator we can give a semantics for in $\mathcal{T}$, we can replicate that operator's logic in an extension of $\mathcal{T}'$, but this does not tell us anything about the relationship between the complexity of $\mathcal{T}'$ to the complexity of $\mathcal{T}$. Ideally, we would like to know whether we can replicate all operators from possible extensions of $\mathcal{T}$ *without an upgrade in computational complexity*.

I believe this work will pay handsome dividends. Potential expressibility results cannot on their own determine which of two semantic theories is correct, but they can clarify the dialectical relationship in which competing theories stand, thus clarifying what kinds of evidence we can expect to find for and against them based on embedding data, and making precise one sense in which a theory can be simpler than another.

[42] See Kamp 1970; Cresswell 1990.
[43] See Peters and Westerståhl (2008, Parts 3-4) for an overview.

A Definitions and proofs

In this appendix I give definitions of the technical terms used in §3, and provide proofs of the claims made there.

A.1 Definitions

Definition A.1. *Languages, Models, Semantic theories*: Given a propositional language $\mathcal{L}$, built from a vocabulary comprising a set $\mathcal{A}$ of atomic sentences $p, q, r \ldots$ and sentence operators $\mathcal{O}$, and comprising only (and typically all) (i) atoms from $\mathcal{A}$, and (ii) strings of the form $O^n(\langle \varphi_1, \varphi_2, \ldots \varphi_n \rangle)$ for any n-place sentence operator $O^n \in \mathcal{O}$ and sentences $\varphi_i : 1 \leq i \leq n$ in $\mathcal{L}$,[44] a *model* $\mathcal{M}$ of $\mathcal{L}$ is a sequence $\langle \mathcal{W}, v, \mathfrak{I}, [\![\cdot]\!] \rangle$, where $\mathcal{W}$ is a non-empty set of possible worlds (any set); v is an atomic valuation function, which takes any atomic sentence of $\mathcal{L}$ and any possible world from $\mathcal{W}$ to either 1 ("true") or 0 ("false"), and which takes any atomic sentence of $\mathcal{L}$ to the subset of $\mathcal{W}$ where that sentence is true; and $\mathfrak{I}$ is a non-empty set of indices (again, any set). $[\![\cdot]\!]$ is an interpretation function for $\mathcal{L}$ which takes an atomic sentence to a set of indices in the model; takes an n-place sentence operator O^n to a function from an n-tuple of sets of indices to a set of indices; take any sentence of the form $O^n(\langle \varphi_1, \varphi_2 \ldots \varphi_n \rangle)$, for any n-place sentence operator $O^n \in \mathcal{O}$ and any n-tuple $\langle \varphi_1, \varphi_2, \ldots \varphi_n \rangle$ of sentences of $\mathcal{L}$, to $[\![O^n]\!](\langle [\![\varphi_1]\!], [\![\varphi_2]\!] \ldots [\![\varphi_n]\!] \rangle)$; and which is otherwise undefined. For convenience, we also stipulate that $[\![\cdot]\!]$ takes any sentence φ of $\mathcal{L}$ and any index i (written $[\![\varphi]\!]^i$) to 1 just in case $i \in [\![\varphi]\!]$, and otherwise to 0. Finally, a *semantic theory* $\mathcal{T}$ of $\mathcal{L}$ is a non-empty set of models of $\mathcal{L}$.

Given a quantified language $\mathcal{L}_v$, built from a vocabulary comprising a set $\mathcal{V}$ of variables $x_1, x_2, \ldots$, a set $\mathfrak{R}$ of relation symbols, and a set $\mathcal{O}$ of sentence operators; and comprising only (and typically all) (i) atoms of the form $R^n(\langle x_1, x_2, \ldots x_n \rangle)$, for R^n an n-place relation in $\mathfrak{R}$, and $x_i : 1 \leq i \leq n$ variables from $\mathcal{V}$; and (ii) strings of the form $O^n(\langle \varphi_1, \varphi_2, \ldots \varphi_n \rangle)$ for any n-place sentence operator $O^n \in \mathcal{O}$ and any sentences $\varphi_i : 1 \leq i \leq n$ in $\mathcal{L}_v$, a *model of* $\mathcal{L}_v$ is just like a model of a propositional language, except it also includes a domain D of individuals, and the atomic valuation function takes a possible world and an n-place relation symbol to an n-place relation (a subset of D^n, the set of n-tuples of elements of D). Interpretation functions and semantic theories are constructed as for propositional languages.

Definition A.2. *Extension of a Language*: Given a language $\mathcal{L}$, and a set $\mathcal{O}$ of sentence operators disjoint from the vocabulary of $\mathcal{L}$,[45] the extension of $\mathcal{L}$ to $\mathcal{O}$, written $\mathcal{L}^{\mathcal{O}}$, is the smallest set containing $\mathcal{L}$ and closed under the elements of $\mathcal{O}$, i.e. where $\mathfrak{a}$ is the function giving the arity of sentence operators, the smallest set containing $\mathcal{L}$ such that if $O \in \mathcal{O}$, $\mathfrak{a}(O) = n$, and $\mathfrak{l} \in (\mathcal{L}^{\mathcal{O}})^n$, then $O^n(\mathfrak{l}) \in \mathcal{L}^{\mathcal{O}}$.

Definition A.3. *Extension of a model and semantic theory*: Given a language $\mathcal{L}$, a model $\mathcal{M}$ of $\mathcal{L}$, and an extension $\mathcal{L}^{\mathcal{O}}$ of $\mathcal{L}$, an extension $\mathcal{M}^{\mathcal{O}}$ of $\mathcal{M}$ to an interpretation of $\mathcal{L}^{\mathcal{O}}$ with respect to $\mathcal{L}$ is an interpretation which is exactly like $\mathcal{M}$ except with respect to its interpretation function, which must agree with $\mathcal{M}$'s interpretation function on sentences of $\mathcal{L}$, i.e. $\forall \varphi \in \mathcal{L} : [\![\varphi]\!]_{\mathcal{M}} \subseteq [\![\varphi]\!]_{\mathcal{M}^{\mathcal{O}}}$. Given a semantic theory $\mathcal{T}$ of $\mathcal{L}$, an extension $\mathcal{T}^{\mathcal{O}}$ of $\mathcal{T}$ to $\mathcal{L}^{\mathcal{O}}$ is a semantic theory each of whose members extends some model in $\mathcal{T}$ from $\mathcal{L}$ to $\mathcal{L}^{\mathcal{O}}$ such that any two models in $\mathcal{T}^{\mathcal{O}}$ agree on the logic of $\mathcal{O}$.[46]

[44] We will generally use lower-case italic letters to range over atoms, and Greek letters to range over all sentences.

[45] I will call any set of operators which meets this novelty constraint a 'set of new operators'; I sometimes leave this novelty condition implicit for brevity in introducing extensions of languages.

[46] I will usually leave the relativization to the initial language implicit. For any model $\mathcal{M}$, I write $[\![\cdot]\!]_{\mathcal{M}}$ for $\mathcal{M}$'s interpretation function, and likewise for its other parameters.

A.2 Proofs

For convenience, I repeat the definition of relative potential expressibility here; I then turn to proofs of the claims of §3:

Definition 3.1. *Relative Potential Expressibility*: For any semantic theories $\mathcal{T}$ and $\mathcal{T}'$ of a language $\mathcal{L}$, $\mathcal{T} \preceq_{\mathcal{L}} \mathcal{T}'$ iff, for any set of new operators $\mathcal{O}$, for any extension $\mathcal{T}^{\mathcal{O}}$ of $\mathcal{T}$ to $\mathcal{L}^{\mathcal{O}}$, there is an extension $\mathcal{T}'^{\mathcal{O}}$ of $\mathcal{T}'$ to $\mathcal{L}^{\mathcal{O}}$ which agrees with $\mathcal{T}^{\mathcal{O}}$ on logic of $\mathcal{O}$: that is, which is such that, for any $\Gamma \subseteq \mathcal{L}^{\mathcal{O}}$ such that $\exists \varphi \in \Gamma : \varphi \in \mathcal{L}^{\mathcal{O}} \setminus \mathcal{L}$, and for any sentence ψ in $\mathcal{L}^{\mathcal{O}}$, $(\Gamma \vDash_{\mathcal{T}^{\mathcal{O}}} \psi) \leftrightarrow (\Gamma \vDash_{\mathcal{T}'^{\mathcal{O}}} \psi)$. For a set of sentences Γ, sentence ψ, and semantic theory $\mathcal{T}$, $\Gamma \vDash_{\mathcal{T}} \psi$ iff for every model $\mathcal{M} \in \mathcal{T}$, $\Gamma \vDash_{\mathcal{M}} \psi$ iff every index in $\mathcal{M}$ which makes all the sentences in Γ true also makes ψ true.

We also define a derivative notion of relative expressibility between models: for any models $\mathcal{M}$ and $\mathcal{M}'$ of a language $\mathcal{L}$, $\mathcal{M} \preceq_{\mathcal{L}} \mathcal{M}'$ iff, for any set of new operators $\mathcal{O}$, for any extension $\mathcal{M}^{\mathcal{O}}$ of $\mathcal{M}$ to $\mathcal{L}^{\mathcal{O}}$, there is an extension $\mathcal{M}'^{\mathcal{O}}$ of $\mathcal{M}'$ to $\mathcal{L}^{\mathcal{O}}$ which preserves the logic of $\mathcal{O}$ from $\mathcal{M}^{\mathcal{O}}$: that is, which is such that, for any $\Gamma \subseteq \mathcal{L}^{\mathcal{O}}$ such that $\exists \varphi \in \Gamma : \varphi \in \mathcal{L}^{\mathcal{O}} \setminus \mathcal{L}$, and for any sentence ψ in $\mathcal{L}^{\mathcal{O}}$, $(\Gamma \vDash_{\mathcal{M}^{\mathcal{O}}} \psi) \leftrightarrow (\Gamma \vDash_{\mathcal{M}'^{\mathcal{O}}} \psi)$.

The proof of Fact 3.1 goes by way of two lemmas:

Lemma A.1. For any models $\mathcal{M}$ and $\mathcal{M}'$ of a language $\mathcal{L}$, $\mathcal{M} \preceq_{\mathcal{L}} \mathcal{M}'$ iff for any extension $\mathcal{L}^{\mathcal{O}}$ of $\mathcal{L}$ with a set of new operators, and any extension $\mathcal{M}^{\mathcal{O}}$ of $\mathcal{M}$ to a model of $\mathcal{L}^{\mathcal{O}}$, there is an extension $\mathcal{M}'^{\mathcal{O}}$ of $\mathcal{M}'$ to $\mathcal{L}^{\mathcal{O}}$ such that there exists a function g from the indices of $\mathcal{M}$ to the indices of $\mathcal{M}'$ such that for any sentence φ of $\mathcal{L}^{\mathcal{O}}$ and index i in $\mathcal{M}$, φ is true at i in $\mathcal{M}^{\mathcal{O}}$ iff φ is true at $g(i)$ in $\mathcal{M}'^{\mathcal{O}}$.

Proof. [⇒] For arbitrary models $\mathcal{M}$ and $\mathcal{M}'$ of an arbitrary language $\mathcal{L}$, suppose $\mathcal{M} \preceq_{\mathcal{L}} \mathcal{M}'$. Recall that means that, for any set of new operators $\mathcal{O}$ and any extension $\mathcal{M}^{\mathcal{O}}$ of $\mathcal{M}$ to $\mathcal{L}^{\mathcal{O}}$, there is an extension $\mathcal{M}'^{\mathcal{O}}$ of $\mathcal{M}'$ to $\mathcal{L}^{\mathcal{O}}$ which preserves the logic of $\mathcal{O}$ from $\mathcal{M}^{\mathcal{O}}$. Consider an arbitrary set of new operators $\mathcal{O}$ and arbitrary extension $\mathcal{M}^{\mathcal{O}}$ of $\mathcal{M}$ to $\mathcal{L}^{\mathcal{O}}$. We will show that there is an extension $\mathcal{M}'^{\mathcal{O}}_{\mathcal{N}}$ of $\mathcal{M}'$ to $\mathcal{L}^{\mathcal{O}}$ such that there is a function g from the indices of $\mathcal{M}$ to the indices of $\mathcal{M}'$ such that for any sentence φ of $\mathcal{L}^{\mathcal{O}}$ and index i in $\mathcal{M}$, φ is true at i in $\mathcal{M}^{\mathcal{O}}$ iff φ is true at $g(i)$ in $\mathcal{M}'^{\mathcal{O}}_{\mathcal{N}}$.

We do so by way of considering first a different extended language which contains $\mathcal{L}^{\mathcal{O}}$, and a different extension of $\mathcal{M}$ which also extends $\mathcal{M}^{\mathcal{O}}$. In particular, consider the extension of $\mathcal{L}$ to $\mathcal{O} \cup \mathcal{N}$, where $\mathcal{O} \cap \mathcal{N} = \varnothing$, no operator in $\mathcal{N}$ is in the vocabulary of $\mathcal{L}$, and the cardinality of $\mathcal{N}$ is greater than the cardinality of the set of indices of $\mathcal{M}$ (if that set is finite) or equal in cardinality to the set of indices of $\mathcal{M}$ (if that set is infinite). Now extend $\mathcal{M}$ to a new model $\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}$ of the resulting language, $\mathcal{L}^{\mathcal{O} \cup \mathcal{N}}$, with the following properties:

(a) $\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}$ is an extension of $\mathcal{M}^{\mathcal{O}}$, so that $\forall \varphi \in \mathcal{L}^{\mathcal{O}} : [\![\varphi]\!]_{\mathcal{M}^{\mathcal{O}}} = [\![\varphi]\!]_{\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}}$, and the indices of $\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}$ are the same as the indices of $\mathcal{M}^{\mathcal{O}}$;

(b) for some sentence $\psi \in \mathcal{L}^{\mathcal{O} \cup \mathcal{N}}$, for each index i of $\mathcal{M}$, there is an $O \in \mathcal{N}$, call it O_i, which uniquely specifies i, in the sense that $O_i(\psi)$ is true at i in $\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}$ and false everywhere else in $\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}$; and

(c) for some unary sentence operator $\neg \in \mathcal{N}$, $\neg$ is given the classical semantics of negation in $\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}$, i.e. for any $\varphi \in \mathcal{L}^{\mathcal{O} \cup \mathcal{N}}$, $\neg(\varphi)$ is true$_{\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}}$ at i iff φ is not true$_{\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}}$ at i.

Now extend $\mathcal{M}'$ to a model $\mathcal{M}'^{\mathcal{O}}_{\mathcal{N}}$ of $\mathcal{L}^{\mathcal{O} \cup \mathcal{N}}$ which preserves the logic of $\mathcal{O} \cup \mathcal{N}$ from $\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}$; we know this will be possible by our assumption that $\mathcal{M} \preceq_{\mathcal{L}} \mathcal{M}'$. Now, define a function g such that, for any index i in $\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}$, g takes i to an index $g(i)$ in $\mathcal{M}'^{\mathcal{O}}_{\mathcal{N}}$ such that (i) $O_i(\psi)$ is true at $g(i)$ in $\mathcal{M}'^{\mathcal{O}}_{\mathcal{N}}$ and (ii) some sentence in $\mathcal{L}^{\mathcal{O} \cup \mathcal{N}}$ is false at $g(i)$. We know there is such an index; otherwise, we would have $O_i(\psi) \vDash_{\mathcal{M}'^{\mathcal{O}}_{\mathcal{N}}} \neg(O_i(\psi))$, and thus $O_i(\psi) \vDash_{\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}} \neg(O_i(\psi))$, but we know the latter is false by our semantics for O_i and $\neg$ in $\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}$.

Now for any sentence φ of $\mathcal{L}^{\mathcal{O} \cup \mathcal{N}}$:

- Suppose first φ is true at i in $\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}$. Then $O_i(\psi) \models_{\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}} \varphi$, and thus by our assumption that O_i has the same logic in $\mathcal{M}'^{\mathcal{O}}_{\mathcal{N}}$ as in $\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}$, $O_i(\psi) \models_{\mathcal{M}'^{\mathcal{O}}_{\mathcal{N}}} \varphi$, and thus φ is true at $g(i)$, since $O_i(\psi)$ is true at $g(i)$.
- Suppose next that φ is not true at i in $\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}$. Then $\neg\varphi$ is true at i in $\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}$, and thus $O_i(\psi) \models_{\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}} \neg\varphi$ and thus $O_i(\psi) \models_{\mathcal{M}'^{\mathcal{O}}_{\mathcal{N}}} \neg\varphi$, and thus $\neg\varphi$ is true at $g(i)$ in $\mathcal{M}'^{\mathcal{O}}_{\mathcal{N}}$, since $O_i(\psi)$ is true at $g(i)$. Thus we can conclude that φ is not true at $g(i)$ in $\mathcal{M}'^{\mathcal{O}}_{\mathcal{N}}$; if it were, since our logic for negation is classical in $\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}$, and thus in $\mathcal{M}'^{\mathcal{O}}_{\mathcal{N}}$, we would have that everything is true at $g(i)$ in $\mathcal{M}'^{\mathcal{O}}_{\mathcal{N}}$, contrary to assumption.

Thus for any φ in $\mathcal{L}^{\mathcal{O}\cup\mathcal{N}}$, we have φ true at i in $\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}$ iff φ is true at $g(i)$ in $\mathcal{M}'^{\mathcal{O}}_{\mathcal{N}}$; thus in particular, for any φ in $\mathcal{L}^{\mathcal{O}}$, which is a subset of $\mathcal{L}^{\mathcal{O}\cup\mathcal{N}}$, φ is true at i in $\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}$ iff φ is true at $g(i)$ in $\mathcal{M}'^{\mathcal{O}}_{\mathcal{N}}$; and, since $\mathcal{M}^{\mathcal{O}}_{\mathcal{N}}$ is an extension of $\mathcal{M}^{\mathcal{O}}$, it follows that for any φ in $\mathcal{L}^{\mathcal{O}}$, φ is true at i in $\mathcal{M}^{\mathcal{O}}$ iff φ is true at $g(i)$ in $\mathcal{M}'^{\mathcal{O}}_{\mathcal{N}}$. Since $\mathcal{O}$ and $\mathcal{M}^{\mathcal{O}}$ were selected arbitrarily, this shows that, for any extension $\mathcal{L}^{\mathcal{O}}$ of $\mathcal{L}$ and extension $\mathcal{M}^{\mathcal{O}}$ of $\mathcal{M}$, we can find an extension of $\mathcal{M}'$ to $\mathcal{L}^{\mathcal{O}}$ with the property that there is a function from the indices of $\mathcal{M}$ to those of $\mathcal{M}'$ which preserves truth and falsity for the sentences of $\mathcal{L}^{\mathcal{O}}$ in the extended models.

[$\Leftarrow$] For arbitrary models $\mathcal{M}$ and $\mathcal{M}'$ of an arbitrary language $\mathcal{L}$, suppose that, for any arbitrary extension $\mathcal{L}^{\mathcal{O}}$ of $\mathcal{L}$ with a set of sentence operators, and any arbitrary extension $\mathcal{M}^{\mathcal{O}}$ of $\mathcal{M}$ to a model of $\mathcal{L}^{\mathcal{O}}$, there is an extension $\mathcal{M}'^{\mathcal{O}}$ of $\mathcal{M}'$ to $\mathcal{L}^{\mathcal{O}}$ such that there exists a function g from the indices of $\mathcal{M}$ to the indices of $\mathcal{M}'$ such that for any sentence $\varphi \in \mathcal{L}^{\mathcal{O}}$ and index i in $\mathcal{M}$, φ is true at i in $\mathcal{M}^{\mathcal{O}}$ iff φ is true at $g(i)$ in $\mathcal{M}'^{\mathcal{O}}$. We can use this fact to construct a new extension $\mathcal{M}'^{\mathcal{O}-}$ of $\mathcal{M}'$ which matches the logic of $\mathcal{O}$ in $\mathcal{M}^{\mathcal{O}}$, as follows. Let $\mathcal{M}'^{\mathcal{O}-}$ be just like $\mathcal{M}'^{\mathcal{O}}$, except that, at every index i of $\mathcal{M}'^{\mathcal{O}}$ which is not in the image of g, for any sentence $\varphi \in \mathcal{L}^{\mathcal{O}} \setminus \mathcal{L}$, φ is false at i in $\mathcal{M}'^{\mathcal{O}-}$ (we can do this because the truth of φ at i will depend just on the semantics we give to our new operators, since if φ is in $\mathcal{L}^{\mathcal{O}} \setminus \mathcal{L}$, it must by definition of $\mathcal{L}^{\mathcal{O}}$ have an operator from $\mathcal{O}$ with highest scope in the sentence). Note that $\mathcal{M}'^{\mathcal{O}-}$ is still an extension of $\mathcal{M}'$ to $\mathcal{L}^{\mathcal{O}}$; and g will still preserve truth for the relevant sentences: since we did not change the truth of any sentences in the image of g, for any $\varphi \in \mathcal{L}^{\mathcal{O}}$, φ is true$_{\mathcal{M}^{\mathcal{O}}}$ iff φ is true$_{\mathcal{M}'^{\mathcal{O}-}}$ at $g(i)$.[47] Now suppose that $\Gamma \models_{\mathcal{M}^{\mathcal{O}}} \psi$ for $(\Gamma \cup \{\psi\}) \subseteq \mathcal{L}^{\mathcal{O}}$, and that $\exists\varphi \in \Gamma : \varphi \in \mathcal{L}^{\mathcal{O}} \setminus \mathcal{L}$. Then by the fact that g preserves truth for sentences of $\mathcal{L}^{\mathcal{O}}$ between $\mathcal{M}^{\mathcal{O}}$ and $\mathcal{M}'^{\mathcal{O}-}$, we have that, within the image of g, ψ is true$_{\mathcal{M}'^{\mathcal{O}-}}$ everywhere that all the members of Γ are; and by construction we have that one member of Γ, namely φ, is false$_{\mathcal{M}'^{\mathcal{O}-}}$ everywhere outside of the image of g; and so $\Gamma \models_{\mathcal{M}'^{\mathcal{O}-}} \psi$. Likewise suppose that $\Gamma \not\models_{\mathcal{M}^{\mathcal{O}}} \psi$; then there is some i where all of Γ is true$_{\mathcal{M}^{\mathcal{O}}}$ and ψ is false$_{\mathcal{M}^{\mathcal{O}}}$, and so at $g(i)$ all of Γ is true$_{\mathcal{M}'^{\mathcal{O}-}}$ with ψ is false$_{\mathcal{M}'^{\mathcal{O}-}}$, and thus we have $\Gamma \not\models_{\mathcal{M}'^{\mathcal{O}-}} \psi$. Thus we have $(\Gamma \models_{\mathcal{M}^{\mathcal{O}}} \psi) \leftrightarrow (\Gamma \models_{\mathcal{M}'^{\mathcal{O}-}} \psi)$. Since this construction was perfectly general, it shows that, under our assumption, for any extension $\mathcal{M}^{\mathcal{O}}$ of $\mathcal{M}$ to an extension $\mathcal{L}^{\mathcal{O}}$ of $\mathcal{L}$, we can find an extension of $\mathcal{M}'$ to $\mathcal{L}^{\mathcal{O}}$ which matches the logic of $\mathcal{O}$ in $\mathcal{M}^{\mathcal{O}}$, and so $\mathcal{M} \preceq_{\mathcal{L}} \mathcal{M}'$.

□

We turn to our second lemma:

Lemma A.2. *Characterization of Model Expressibility*: For any models $\mathcal{M}$ and $\mathcal{M}'$ and language $\mathcal{L}$, $\mathcal{M} \preceq_{\mathcal{L}} \mathcal{M}'$ iff there is a function g (call it a *witness function* with respect to $\mathcal{L}$) from the indices of $\mathcal{M}$ to those of $\mathcal{M}'$ which is such that (i) for any sentence φ of $\mathcal{L}$ and index i in $\mathcal{M}$, φ is true at i in $\mathcal{M}$ iff φ is true at $g(i)$ in $\mathcal{M}'$; and (ii) g is an injection.

Proof. [$\Rightarrow$] Suppose for arbitrary $\mathcal{M}$, $\mathcal{M}'$ and $\mathcal{L}$, there is no function g from the indices of $\mathcal{M}$ to those of $\mathcal{M}'$ which is such that (i) for any sentence φ of $\mathcal{L}$ and index i in $\mathcal{M}$, φ is true at i in $\mathcal{M}$ iff φ is true at $g(i)$ in $\mathcal{M}'$; and (ii) g is an injection. Find a set of new operators $\mathcal{O}$ with cardinality equal to the set of indices in $\mathcal{M}$. Let f be a bijection from the indices of $\mathcal{M}$ to $\mathcal{O}$. Extend $\mathcal{M}$ to a new model $\mathcal{M}^{\mathcal{O}}$ of $\mathcal{L}^{\mathcal{O}}$, with the property that, for any index i in $\mathcal{M}$, for any sentence $\varphi \in \mathcal{L}^{\mathcal{O}}$, $f(i)(\varphi)$ is true$_{\mathcal{M}^{\mathcal{O}}}$ at i and false$_{\mathcal{M}^{\mathcal{O}}}$ at every other index of

[47] 'True$_{\mathcal{M}}$' is shorthand for 'true in $\mathcal{M}$'.

$\mathcal{M}$; that is, $f(i)$ "tags" i in $\mathcal{M}^{\mathcal{O}}$. Now consider an arbitrary extension $\mathcal{M}'^{\mathcal{O}}$ of $\mathcal{M}'$. Suppose there is a function g from the indices of $\mathcal{M}$ to the indices of $\mathcal{M}'$ with the property that, for all $\varphi \in \mathcal{L}^{\mathcal{O}}$, φ is true$_{\mathcal{M}^{\mathcal{O}}}$ at i iff φ is true$_{\mathcal{M}'^{\mathcal{O}}}$ at $g(i)$. Since $\mathcal{L} \subset \mathcal{L}^{\mathcal{O}}$, and since extensions of models of a given language preserve truth for sentences in the original language, we know that, for all $\varphi \in \mathcal{L}$, φ is true$_{\mathcal{M}}$ at i iff φ is true$_{\mathcal{M}'}$ at $g(i)$. Then it follows from our assumption that g is not an injection: for some $\mathcal{M}$-indices i and i' with $i \neq i'$, $g(i) = g(i')$. Choose some $\varphi \in \mathcal{L}^{\mathcal{O}}$. We know by construction of $\mathcal{M}^{\mathcal{O}}$ that $f(i)(\varphi)$ is true$_{\mathcal{M}^{\mathcal{O}}}$ at i and false$_{\mathcal{M}^{\mathcal{O}}}$ at i'. But, since $g(i) = g(i')$, $f(i)(\varphi)$ will either be true$_{\mathcal{M}'^{\mathcal{O}}}$ at both $g(i)$ and $g(i')$, or false at both, and thus it will not be the case that, for every sentence φ of $\mathcal{L}^{\mathcal{O}}$, if φ is true at i in $\mathcal{M}^{\mathcal{O}}$, then φ is true at $g(i)$ in $\mathcal{M}'^{\mathcal{O}}$, contrary to assumption. Thus, since $\mathcal{M}'^{\mathcal{O}}$ was chosen arbitrarily, there is no extension of $\mathcal{M}'$ to $\mathcal{L}^{\mathcal{O}}$ such that there is a function which preserves truth and falsity for all sentences in $\mathcal{L}^{\mathcal{O}}$ between $\mathcal{M}^{\mathcal{O}}$ and $\mathcal{M}'^{\mathcal{O}}$; and thus by Lemma A.1, $\mathcal{M} \not\preceq_{\mathcal{L}} \mathcal{M}'$.

[$\Leftarrow$] Suppose, for arbitrary $\mathcal{M}$, $\mathcal{M}'$ and $\mathcal{L}$, there is such a truth-preserving injection g. Given an arbitrary set $\mathcal{O}$ of sentence operators and an arbitrary extension $\mathcal{M}^{\mathcal{O}}$ of $\mathcal{M}$ to $\mathcal{L}^{\mathcal{O}}$, we show there is an extension $\mathcal{M}'^{\mathcal{O}}$ of $\mathcal{M}'$ to $\mathcal{L}^{\mathcal{O}}$ which has the property that, for any sentence $\varphi \in \mathcal{L}^{\mathcal{O}}$, φ is true at an index i in $\mathcal{M}^{\mathcal{O}}$ just in case φ is true at $g(i)$ in $\mathcal{M}'^{\mathcal{O}}$. Let K index the elements of $\mathcal{O}$. For each $O_k : k \in K$, extend $\mathcal{M}$ to the model $\mathcal{M}_k$ which is just like $\mathcal{M}$, except its interpretation function $[\![\cdot]\!]_{\mathcal{M}_k}$ is augmented with the semantic rule for O_k from $[\![\cdot]\!]_{\mathcal{M}^{\mathcal{O}}}$. Then, for each O_k, extend $\mathcal{M}'$ to the model $\mathcal{M}'_k$ which augments the interpretation function of $\mathcal{M}'$ with a semantic rule for O_k as follows. For brevity, for any set α and function f, define $f[\alpha]$ to be the pointwise application of f to α where defined, i.e. $f[\alpha] = \{f(a) : a \in \alpha \wedge f(a) \text{ is defined}\}$. Let g^{-1} be the inverse of g, defined only on the image of g; that g^{-1} is a well-defined function follows because g is an injection. Now, suppose first that O_k is a unary sentence operator; then let $\mathcal{M}'_k$ extend $\mathcal{M}'$ with the following semantic rule: $[\![O_k]\!]_{\mathcal{M}'_k} = \lambda s_{\mathcal{M}'}.g[[\![O_k]\!]_{\mathcal{M}_k}(g^{-1}[s])]$, where $s_{\mathcal{M}'}$ ranges over sets of $\mathcal{M}'$ indices. Thus in $\mathcal{M}'_k$, O_k takes a set of $\mathcal{M}'$ indices; then finds the pre-image (where defined) of this complement with respect to g; then applies the semantic rule for O_k in $\mathcal{M}_k$ to this pre-image; and finally, returns the pointwise application of g to the resulting set. Now note that, for any set s of $\mathcal{M}$-indices and set s' of $\mathcal{M}'$-indices, if $i \in s \leftrightarrow g(i) \in s'$, it follows that $i \in [\![O_k]\!]_{\mathcal{M}_k}(s) \leftrightarrow g(i) \in [\![O_k]\!]_{\mathcal{M}'_k}(s')$. To see this, assume for arbitrary s and s' that $i \in s \leftrightarrow g(i) \in s'$. Now note that $s = g^{-1}[s']$: if $i \in s$, then by assumption $g(i) \in s'$, and thus $g^{-1}[s']$ will include i, by construction; and if $i \notin s$, then by assumption $g(i) \notin s'$, and since g^{-1} is an injection, by construction, we know that $i \notin g^{-1}[s']$. We thus have $[\![O_k]\!]_{\mathcal{M}'_k}(s') = g[[\![O_k]\!]_{\mathcal{M}_k}(g^{-1}[s'])] = g[[\![O_k]\!]_{\mathcal{M}_k}(s)]$. In other words, whenever $i \in s \leftrightarrow g(i) \in s'$, then $[\![O_k]\!]_{\mathcal{M}'_k}(s')$ is just the pointwise application of g to $[\![O_k]\!]_{\mathcal{M}_k}(s)$, and thus, since g is an injection, $i \in [\![O_k]\!]_{\mathcal{M}_k}(s) \leftrightarrow g(i) \in [\![O_k]\!]_{\mathcal{M}'_k}(s')$. The generalization of this construction to n-place sentence operators, for any n, is straightforward. We use this method to construct $\mathcal{M}'_k$ for all $k \in K$.

Now, where $\mathcal{M}' = \langle \mathcal{W}, \mathfrak{I}, v, [\![\cdot]\!]_{\mathcal{M}'} \rangle$ or $\langle D, \mathcal{W}, \mathfrak{I}, v, [\![\cdot]\!]_{\mathcal{M}'} \rangle$, let $\mathcal{M}'^{\mathcal{O}} = \left\langle \mathcal{W}, \mathfrak{I}, v, \bigcup_{k \in K} [\![\cdot]\!]_{\mathcal{M}'_k} \right\rangle$ or $\left\langle D, \mathcal{W}, \mathfrak{I}, v, \bigcup_{k \in K} [\![\cdot]\!]_{\mathcal{M}'_k} \right\rangle$, respectively. By our construction, we know that for any $O \in \mathcal{O}$ and any sets of $\mathcal{M}$-indices s and $\mathcal{M}'$-indices s' such that $i \in s \leftrightarrow g(i) \in s'$, $[\![O]\!]_{\mathcal{M}'^{\mathcal{O}}}(s') = g[[\![O]\!]_{\mathcal{M}^{\mathcal{O}}}(s)]$, and thus $i \in [\![O]\!]_{\mathcal{M}^{\mathcal{O}}}(s) \leftrightarrow g(i) \in [\![O]\!]_{\mathcal{M}'^{\mathcal{O}}}(s')$. We know by assumption that, for all sentences $\varphi \in \mathcal{L}$, $i \in [\![\varphi]\!]_{\mathcal{M}} \leftrightarrow g(i) \in [\![\varphi]\!]_{\mathcal{M}'}$, and thus (since extending a model never changes its interpretation of a sentence already in the language of the original model) $i \in [\![\varphi]\!]_{\mathcal{M}^{\mathcal{O}}} \leftrightarrow g(i) \in [\![\varphi]\!]_{\mathcal{M}'^{\mathcal{O}}}$. Now consider any sequence of sentences $\vec{\psi}$ with the property that for each ψ_j in the sequence, $i \in [\![\psi_j]\!]_{\mathcal{M}^{\mathcal{O}}} \leftrightarrow g(i) \in [\![\psi_j]\!]_{\mathcal{M}'^{\mathcal{O}}}$. Then we know that, by our construction, for any $k \in K$ and index i of $\mathcal{M}$, $i \in [\![O_k(\vec{\psi})]\!]_{\mathcal{M}^{\mathcal{O}}} \leftrightarrow g(i) \in [\![O_k(\vec{\psi})]\!]_{\mathcal{M}'^{\mathcal{O}}}$. Since the sentences of $\mathcal{L}^{\mathcal{O}}$ are built recursively from the sentences of $\mathcal{L}$ and the operators in $\mathcal{O}$, it follows by an induction on the complexity of formulae that, for any $\varphi \in \mathcal{L}^{\mathcal{O}}$, $i \in [\![\varphi]\!]_{\mathcal{M}^{\mathcal{O}}} \leftrightarrow g(i) \in [\![\varphi]\!]_{\mathcal{M}'^{\mathcal{O}}}$. Since $\mathcal{O}$ and $\mathcal{M}^{\mathcal{O}}$ were chosen arbitrarily, we conclude that, for any set of new operators $\mathcal{O}$, for any extension $\mathcal{M}^{\mathcal{O}}$ of $\mathcal{M}$ to $\mathcal{L}^{\mathcal{O}}$, there is an extension $\mathcal{M}'^{\mathcal{O}}$ of $\mathcal{M}'$ to $\mathcal{L}^{\mathcal{O}}$ such that there is a function g with the property that, for any sentence $\varphi \in \mathcal{L}^{\mathcal{O}}$, φ

is true at an index i in $\mathcal{M}^{\mathcal{O}}$ just in case φ is true at $g(i)$ in $\mathcal{M}'^{\mathcal{O}}$; and thus by Lemma A.1, $\mathcal{M} \preceq_{\mathcal{L}} \mathcal{M}'$.

□

We turn now to our proof of Fact 3.1:

Fact 3.1. *Characterization of Expressibility*: For any semantic theories $\mathcal{T}$ and $\mathcal{T}'$ and language $\mathcal{L}$, if $\mathcal{T}$ is isomorphic with respect to $\mathcal{L}$, and $\mathcal{T}'$ is isomorphic with respect to $\mathcal{L}$, then $\mathcal{T} \preceq_{\mathcal{L}} \mathcal{T}'$ iff there is a model $\mathcal{M} \in \mathcal{T}$ and a model $\mathcal{M}' \in \mathcal{T}'$ such that there is a witness function g from the indices of $\mathcal{M}$ to those of $\mathcal{M}'$ with respect to $\mathcal{L}$. A semantic theory $\mathcal{T}$ is *isomorphic* with respect to $\mathcal{L}$ iff $\forall \mathcal{M}, \mathcal{M}' \in \mathcal{T} : \mathcal{M} \preceq_{\mathcal{L}} \mathcal{M}' \wedge \mathcal{M}' \preceq_{\mathcal{L}} \mathcal{M}$.

Proof. For abitrary semantic theories $\mathcal{T}, \mathcal{T}'$, and language $\mathcal{L}$, suppose that $\mathcal{T}$ and $\mathcal{T}'$ are both isomorphic with respect to $\mathcal{L}$:

[$\Rightarrow$]: Suppose there is no pair of models $\mathcal{M} \in \mathcal{T}$ and $\mathcal{M}' \in \mathcal{T}'$ such that there is a witness function g from the indices of $\mathcal{M}$ to those of $\mathcal{M}'$. It follows by Lemma A.2 that for any models $\mathcal{M} \in \mathcal{T}$ and $\mathcal{M}' \in \mathcal{T}'$, $\mathcal{M} \not\preceq_{\mathcal{L}} \mathcal{M}'$. Choose arbitrary model $\mathcal{M} \in \mathcal{T}$ and $\mathcal{M}' \in \mathcal{T}'$, and find a set of operators $\mathcal{O}$ and extension $\mathcal{M}^{\mathcal{O}}$ of $\mathcal{M}$ to $\mathcal{L}^{\mathcal{O}}$ such that there is no extension of $\mathcal{M}'$ which matches the logic of $\mathcal{O}$ in $\mathcal{M}^{\mathcal{O}}$. We know this will be possible since otherwise $\mathcal{M} \preceq_{\mathcal{L}} \mathcal{M}'$. We can show moreover that no model $\mathcal{M}'' \in \mathcal{T}'$ can be extended to match the logic of $\mathcal{O}$ in $\mathcal{M}^{\mathcal{O}}$; else since $\mathcal{T}'$ is isomorphic, we could extend $\mathcal{M}'$ to match the logic of $\mathcal{O}$ in $\mathcal{M}^{\mathcal{O}}$, contrary to assumption. Consider any extension $\mathcal{T}^{\mathcal{O}}$ of $\mathcal{T}$ to $\mathcal{L}^{\mathcal{O}}$ which includes $\mathcal{M}^{\mathcal{O}}$ and any extension $\mathcal{T}'^{\mathcal{O}}$ of $\mathcal{T}'$ to $\mathcal{L}^{\mathcal{O}}$. If these agreed on the logic of $\mathcal{O}$, then, since all the models within each theory agree with each other on the logic of $\mathcal{O}$, then every model in $\mathcal{T}'^{\mathcal{O}}$ would agree with every model in $\mathcal{T}^{\mathcal{O}}$ on the logic of $\mathcal{O}$, contrary to our assumption that no extension of any model in $\mathcal{T}'$ matches the logic of $\mathcal{O}$ in $\mathcal{M}^{\mathcal{O}}$; so $\mathcal{T}'^{\mathcal{O}}$ does not agree with $\mathcal{T}^{\mathcal{O}}$ on the logic of $\mathcal{O}$; since $\mathcal{T}'^{\mathcal{O}}$ was chosen arbitrarily, it follows that no extension of $\mathcal{T}'$ agrees with $\mathcal{T}^{\mathcal{O}}$ on the logic of $\mathcal{O}$; and so we have $\mathcal{T} \not\preceq_{\mathcal{L}} \mathcal{T}'$.

[$\Leftarrow$] Suppose there is a model $\mathcal{M} \in \mathcal{T}$ and a model $\mathcal{M}' \in \mathcal{T}'$ such that there is a witness function g from the indices of $\mathcal{M}$ to those of $\mathcal{M}'$. It follows by Lemma A.2 that $\mathcal{M} \preceq_{\mathcal{L}} \mathcal{M}'$. Consider any extension $\mathcal{L}^{\mathcal{O}}$ of $\mathcal{L}$ with a new set of operators $\mathcal{O}$, and extension $\mathcal{M}^{\mathcal{O}}$ of $\mathcal{M}$ to a model of $\mathcal{L}^{\mathcal{O}}$. Find an extension $\mathcal{M}'^{\mathcal{O}}$ of $\mathcal{M}'$ of $\mathcal{L}^{\mathcal{O}}$ which matches the logic of $\mathcal{O}$ from $\mathcal{M}^{\mathcal{O}}$. Now extend every model in $\mathcal{T}'$ other than $\mathcal{M}'$ to match the logic of $\mathcal{O}$ in $\mathcal{M}'^{\mathcal{O}}$; we know this will be possible because $\mathcal{T}'$ is isomorphic with respect to $\mathcal{L}$. Call the resulting set of models $\mathcal{T}'^{\mathcal{O}}$. For any way of completing the extension of $\mathcal{T}$ to a new semantic theory $\mathcal{T}^{\mathcal{O}}$ of $\mathcal{L}^{\mathcal{O}}$, all the models in the extension will agree with $\mathcal{M}^{\mathcal{O}}$ on the logic of $\mathcal{O}$, by definition of an extension, and so $\mathcal{T}^{\mathcal{O}}$ and $\mathcal{T}'^{\mathcal{O}}$ will agree on the logic of $\mathcal{O}$. Hence $\mathcal{T} \preceq_{\mathcal{L}} \mathcal{T}'$.

□

Fact 3.2. $\mathcal{D} \preceq_{\mathcal{L}^{\Diamond}} \mathcal{R}$.

Proof. Recall that our language $\mathcal{L}^{\Diamond}$ contains an infinite set of atoms $p, q, r \ldots$ closed under the one-place sentence operator $\Diamond$, and that we assume that all of our semantic theories are sets of models whose sets of worlds and valuation functions are such that, in any model, any two worlds differ on the truth of some atom, and such that for any combination of atoms, exactly that set of atoms is true at some world, according to that model's valuation function. The relational semantics $\mathcal{R}$ is the class of models $\mathfrak{r}$ of the form $\langle \mathcal{W}_{\mathfrak{r}}, v_{\mathfrak{r}}, \mathfrak{I}_{\mathfrak{r}}, [\![\cdot]\!]_{\mathfrak{r}} \rangle$, where $\mathfrak{I}_{\mathfrak{r}} = \{\langle f, w\rangle : f : \mathcal{W}_{\mathfrak{r}} \to \wp(\mathcal{W}_{\mathfrak{r}}) \wedge w \in \mathcal{W}_{\mathfrak{r}}\}$ and $[\![\cdot]\!]_{\mathfrak{r}}$ defined as specifed in §3; likewise the domain semantics $\mathcal{D}$ is the class of models $\mathfrak{d}$ of the form $\langle \mathcal{W}_{\mathfrak{d}}, v_{\mathfrak{d}}, \mathfrak{I}_{\mathfrak{d}}, [\![\cdot]\!]_{\mathfrak{d}} \rangle$, where $\mathfrak{I}_{\mathfrak{d}} = \{\langle s, w\rangle : s \subseteq \mathcal{W}_{\mathfrak{d}} \wedge w \in \mathcal{W}_{\mathfrak{d}}\}$, with the interpretation function again specified as above. $\mathcal{D}$ is clearly isomorphic with respect to $\mathcal{L}^{\Diamond}$, as is $\mathcal{R}$, so by Fact 3.1 it suffices to show that there is a $\mathfrak{d} \in \mathcal{D}$ and an $\mathfrak{r} \in \mathcal{R}$ s.t. $\mathfrak{d} \preceq_{\mathcal{L}^{\Diamond}} \mathfrak{r}$. Choose $\mathfrak{d}$ at random and let

$\mathfrak{r}$ be any model in $\mathcal{R}$ built on the same set of worlds and valuation function as $\mathfrak{d}$. Let g be a function $\mathfrak{I}_\mathfrak{d} \to \mathfrak{I}_\mathfrak{r}$ as follows. For any index $\langle s, w\rangle \in \mathfrak{I}_\mathfrak{d}$, let $g(\langle s, w\rangle) = \langle f^s, w\rangle$, where f^s is the constant function from worlds to s. For any atomic sentence p of $\mathcal{L}^\Diamond$, p will be true$_\mathfrak{d}$ at i iff p is true$_\mathfrak{r}$ at $g(i)$, since we are assuming the same stock of worlds and atomic valuation in both models, and since the truth of atomic sentence in these frameworks depends only on the world parameter of the index and the atomic valuation. Now for any sentence $\varphi \in \mathcal{L}^\Diamond$, assume for induction that φ is true$_\mathfrak{d}$ at i iff it is true$_\mathfrak{r}$ at $g(i)$. We show that, for arbitrary index i, $\Diamond\varphi$ is true$_\mathfrak{d}$ at i iff $\Diamond\varphi$ is true$_\mathfrak{r}$ at $g(i)$. i will have the form $\langle s, w\rangle$, for information state s and world w, and, by our semantics for $\Diamond$ in $\mathfrak{d}$, $\Diamond\varphi$ will be true$_\mathfrak{d}$ at i iff φ is true$_\mathfrak{d}$ at some element in the set $\Phi = \{\langle s, w'\rangle : w' \in s\}$. $g(i)$ will have the form $\langle f^s, w\rangle$, and, by our semantics for $\Diamond$ in $\mathfrak{r}$, $\Diamond\varphi$ will be true$_\mathfrak{r}$ at $g(i)$ iff φ is true$_\mathfrak{r}$ for some element in the set $\Psi = \{\langle f^s, w'\rangle : w' \in f^s(w)\}$. Now note that, thanks to the way we constructed g and the fact that $f^s(w) = s$, g will be a bijection from Φ to Ψ. And so it follows from our assumption for induction that φ will be true$_\mathfrak{d}$ at some element in Φ just in case φ is true$_\mathfrak{r}$ at some element in Ψ, and thus $\Diamond\varphi$ will be true$_\mathfrak{d}$ at i iff $\Diamond\varphi$ is true$_\mathfrak{r}$ at $g(i)$. It thus follows by induction on the complexity of formulae that, for any sentence φ of $\mathcal{L}^\Diamond$ and any index i, φ is true$_\mathfrak{d}$ at i iff φ is true$_\mathfrak{r}$ at $g(i)$. Finally, it is easy to see that g is an injection. Given Fact 3.1, it thus follows that $\mathcal{D} \preceq_{\mathcal{L}^\Diamond} \mathcal{R}$.

□

Fact 3.3. $\mathcal{R} \not\preceq_{\mathcal{L}^\Diamond} \mathcal{D}$.

Proof. Consider any models $\mathfrak{d} \in \mathcal{D}$ and $\mathfrak{r} \in \mathcal{R}$. Consider an $\mathfrak{r}$-index $\langle f, w\rangle$, with $f(w) = \{w'\}$, $v_\mathfrak{r}(p, w') = 0$, $f(w') = \{w''\}$, and $v_\mathfrak{r}(p, w'') = 1$. Then $\Diamond p$ will be false$_\mathfrak{r}$ at $\langle f, w\rangle$, while $\Diamond(\Diamond p)$ will be true$_\mathfrak{r}$ at $\langle f, w\rangle$. There is no function g which replicates this pattern in $\mathfrak{d}$—i.e. which has $\Diamond p$ false$_\mathfrak{d}$ at $g(\langle f, w\rangle)$ and has $\Diamond(\Diamond p)$ true$_\mathfrak{d}$ at $g(\langle f, w\rangle)$. This is for the simple reason that, for any index i in $\mathfrak{I}_\mathfrak{d}$, $\Diamond\varphi$ is true$_\mathfrak{d}$ at i iff $\Diamond(\Diamond\varphi)$ is, since $\Diamond(\Diamond\varphi)$ is true$_\mathfrak{d}$ at $\langle s, x\rangle$, for any x, iff $\Diamond\varphi$ is true$_\mathfrak{d}$ at $\langle s, w'\rangle$ for some $w' \in s$ iff φ is true$_\mathfrak{d}$ at $\langle s, w''\rangle$ for some $w'' \in s$ iff $\Diamond\varphi$ is true$_\mathfrak{d}$ at $\langle s, x\rangle$ for any x. Thus there is no function from the indices of $\mathfrak{r}$ to those of $\mathfrak{d}$ which preserves truth for all $\varphi \in \mathcal{L}^\Diamond$. Since these models were chosen at random, we have that there is no model $\mathfrak{r} \in \mathcal{R}$ and model $\mathfrak{d} \in \mathcal{D}$ s.t. $\mathfrak{r} \preceq_{\mathcal{L}^\Diamond} \mathfrak{d}$ and thus by Fact 3.1 we have $\mathcal{R} \not\preceq_{\mathcal{L}^\Diamond} \mathcal{D}$.

□

Fact 3.5. $\mathcal{D} \prec_{\mathcal{L}^{\Diamond-}} \mathcal{R}$.

Proof. That $\mathcal{D} \preceq_{\mathcal{L}^{\Diamond-}} \mathcal{R}$ follows as an immediate corollary of Fact 3.2. The proof that $\mathcal{R} \not\preceq_{\mathcal{L}^{\Diamond-}} \mathcal{D}$ is as follows. Consider any models $\mathfrak{r} \in \mathcal{R}$ and $\mathfrak{d} \in \mathcal{D}$. Let h be a bijection $\mathcal{W}_\mathfrak{r} \to \mathcal{W}_\mathfrak{d}$ such that $\forall p \in \mathcal{L}^{\Diamond-} : \forall w \in \mathcal{W}_\mathfrak{r} : v_\mathfrak{r}(p, w) = v_\mathfrak{d}(p, h(w))$; that there is such a bijection follows from our assumptions about the stocks of worlds and valuation functions in any model of $\mathcal{D}$ and $\mathcal{R}$. Now consider three different modal bases f, f', and f'' from pairs in $\mathfrak{I}_\mathfrak{r}$, and some world $w \in \mathcal{W}_\mathfrak{r}$, with $f(w) = f'(w) = f''(w) = \varnothing$. Consider any function $g : \mathfrak{I}_\mathfrak{r} \to \mathfrak{I}_\mathfrak{d}$ which preserves truth for $\varphi \in \mathcal{L}^{\Diamond-}$. Suppose that $g(\langle f, w\rangle) = \langle s, x\rangle$, $g(\langle f', w\rangle) = \langle s', x'\rangle$, and $g(\langle f'', w\rangle) = \langle s'', x''\rangle$, with $\langle s, x\rangle, \langle s', x'\rangle$, and $\langle s'', x''\rangle$ all different. Since all worlds differ on the truth of some atom, we know that $x = x' = x'' = h(w)$, else we would have that at least one of $g(\langle f, w\rangle)$, $g(\langle f', w\rangle)$, or $g(\langle f'', w\rangle)$ differs from its pre-image on the truth of some atom, contrary to the assumption that g preserves truth. So we must have that $s \neq s'$ and $s \neq s''$ and $s' \neq s''$. It is easy to see that, for any atom p, $\Diamond p$ is false$_\mathfrak{r}$ at all of $\langle f, w\rangle$, $\langle f', w\rangle$, and $\langle f'', w\rangle$. But there are only two $\mathfrak{d}$-indices with $h(w)$ as their world parameter which make $\Diamond p$ false$_\mathfrak{d}$ for every atom p, namely $\langle \varnothing, h(w)\rangle$ and $\left\langle \{w^f_\mathfrak{d}\}, h(w)\right\rangle$, where $w^f_\mathfrak{d}$ is the world in $\mathcal{W}_\mathfrak{d}$ such that for every atomic sentence p, $v_\mathfrak{d}(p, w^f_\mathfrak{d}) = 0$. And so either $\langle s, x\rangle, \langle s', x'\rangle$, or $\langle s'', x''\rangle$ will make $\Diamond p$ true$_\mathfrak{d}$ for some p, contrary to the assumption that g preserves truth. Thus any truth-preserving function must take two of $\langle f, w\rangle$, $\langle f', w\rangle$, and $\langle f'', w\rangle$ to the same index in $\mathfrak{I}_\mathfrak{d}$, and thus will fail to be an injection. Since $\mathfrak{d}$ and $\mathfrak{r}$ were chosen arbitrarily, Fact 3.5 follows by Fact 3.1.

□

Fact 3.8. $\mathcal{U} \prec_{\mathcal{L}^\Diamond} \mathcal{R}$.

Proof. Recall that $\mathcal{U}$ is the set of models $\mathfrak{u}$ of the form $\langle \mathcal{W}_\mathfrak{u}, v_\mathfrak{u}, \mathfrak{I}_\mathfrak{u}, [\![\cdot]\!]_\mathfrak{u} \rangle$, where $\mathfrak{I}_\mathfrak{u} = \{\langle s, c\rangle : s \cup c \subseteq \mathcal{W}_\mathfrak{u}\}$, with $[\![\cdot]\!]_\mathfrak{u}$ as specified in §3, and assuming again that in any $\mathcal{U}$-model, any two worlds differ on the truth of some atom, and for any set of atoms, there is a world where exactly they are true. First note that $\mathcal{U}$ is isomorphic with respect to $\mathcal{L}^\Diamond$. Next, let $\mathfrak{u}$ be an arbitrary $\mathcal{U}$-model, and let $\mathfrak{r}$ be a model in $\mathcal{R}$ built on the same set of worlds and valuation function. Let $w^t_\mathfrak{u}$ be the world in $\mathcal{W}_\mathfrak{u}$ which is such that, for every atom $p \in \mathcal{L}^\Diamond : v_\mathfrak{u}(p, w^t_\mathfrak{u}) = 1$, and let $w^f_\mathfrak{u}$ be the world in $\mathcal{W}_\mathfrak{u}$ which is such that, for every atom $p \in \mathcal{L}^\Diamond : v_\mathfrak{u}(p, w^f_\mathfrak{u}) = 0$. Where Φ is a set of atomic sentences, we let Φ refer to the unique world from $\mathcal{W}_\mathfrak{u}$ which verifies those sentences according to $v_\mathfrak{u}$, and *vice versa*. Let h be an injection which takes any pair of subsets of $\mathcal{W}_\mathfrak{u}$ to a subset of $\mathcal{W}_\mathfrak{u}$ (that there is such a function follows from the fact that $\mathcal{W}_\mathfrak{u}$ must be infinite given our starting language and assumptions about worlds). We stipulate further that $h(\left\langle \{w^f_\mathfrak{u}\}, \varnothing \right\rangle) = \{w^t_\mathfrak{u}\}$, $h(\langle \{w^t_\mathfrak{u}\}, \{w^t_\mathfrak{u}\}\rangle) = \{w^t_\mathfrak{u}, w^f_\mathfrak{u}\}$; $h(\left\langle \{w^f_\mathfrak{u}, w^t_\mathfrak{u}\}, \{w^t_\mathfrak{u}\} \right\rangle) = \{w^f_\mathfrak{u}\}$; and that $h(\langle s, c\rangle)$ includes $w^t_\mathfrak{u}$ whenever $w^t_\mathfrak{u} \in s$ and $s = c$. For any sets r, s, let f^r_s be the function which takes every world in $\mathcal{W}_\mathfrak{u}$ to r except $w^f_\mathfrak{u}$, which it takes to s, and let f^r_{s*} be the function which takes every world in $\mathcal{W}_\mathfrak{u}$ to r except $w^t_\mathfrak{u}$, which it takes to s. We can then define a witness function g as follows: for any pair $\langle s, c\rangle$ of contexts (subsets of $\mathcal{W}_\mathfrak{u}$), with p ranging over atomic sentences in $\mathcal{L}^\Diamond$:

$$g(\langle s, c\rangle) = \begin{cases} \left\langle f^s_{h(\langle s,c\rangle)*}, \{p : \forall w' \in s : p \in w'\} \right\rangle & \text{iff } s = c \neq \varnothing \\ \left\langle f^\varnothing_{h(\langle s,c\rangle)*}, \{p : (\forall w' \in c : p \in w') \wedge (\forall w'' \in (s \setminus c) : p \notin w'')\} \right\rangle & \text{iff } c \subset s \wedge c \neq \varnothing \\ \left\langle f^\varnothing_{h(\langle s,c\rangle)*}, w^f_\mathfrak{u} \right\rangle & \text{iff } c \not\subseteq s \\ \left\langle f^{\{w : \forall p : p \in w \rightarrow \forall w' \in s : p \notin w'\}}_{h(\langle s,c\rangle)*}, \{p : \forall w' \in s : p \notin w'\} \right\rangle & \text{iff } s \neq c \wedge c = \varnothing \\ \left\langle f^{\{w^t_\mathfrak{u}\}}_{h(\langle s,c\rangle)}, w^t_\mathfrak{u} \right\rangle & \text{iff } s = c = \varnothing \end{cases}$$

First note that g is an injection, since each pair of contexts is taken to an index whose modal base is uniquely tagged by $h(\cdot)$.

Now note that, for any sentence $\varphi \in \mathcal{L}^\Diamond$ and $i \in \mathfrak{I}_\mathfrak{u}$, φ is $\text{true}_\mathfrak{u}$ at i iff φ is $\text{true}_\mathfrak{r}$ at $g(i)$. To see this, consider first atomic q. Atomic q is $\text{true}_\mathfrak{u}$ at $\langle s, c\rangle$ iff c is the result of removing all and only $\bar{q}$-worlds from s:[48]

- if $s = c \neq \varnothing$, then this holds iff all the worlds in s are q-worlds iff $q \in \{p : \forall w' \in s : p \in w'\}$ iff q is $\text{true}_\mathfrak{r}$ at $g(\langle s, c\rangle) = \left\langle f^s_{h(\langle s,c\rangle)*}, \{p : \forall w' \in s : p \in w'\} \right\rangle$;
- if $c \subset s \wedge c \neq \varnothing$, then this holds iff all the worlds in c, but none of the worlds in $s \setminus c$, are q-worlds, iff $q \in \{p : (\forall w' \in c : p \in w') \wedge (\forall w'' \in (s \setminus c) : p \notin w'')\}$, iff q is $\text{true}_\mathfrak{r}$ at $g(\langle s, c\rangle) = \langle f^\varnothing_{h(\langle s,c\rangle)*}, \{p : (\forall w' \in c : p \in w') \wedge (\forall w'' \in s : p \notin w'')\}\rangle$;
- if $c \not\subseteq s$, then this never holds, in which case q is also $\text{false}_\mathfrak{r}$ at $g(\langle s, c\rangle) = \left\langle f^\varnothing_{h(\langle s,c\rangle)*}, w^f_\mathfrak{u} \right\rangle$;
- if $s \neq c \wedge c = \varnothing$, then this holds iff no world in s is a q-world iff $q \in \{p : \forall w' \in s : p \notin w'\}$ iff q is $\text{true}_\mathfrak{r}$ at $g(\langle s, c\rangle) = \langle f^{\{w : \forall p : p \in w \rightarrow \forall w' \in s : p \notin w'\}}_{h(\langle s,c\rangle)*}, \{p : \forall w' \in s : p \notin w'\}\rangle$;
- if $s = c = \varnothing$, this holds in any case whatsoever, in which case q is also $\text{true}_\mathfrak{r}$ at $g(\langle s, c\rangle) = \left\langle f^{\{w^t_\mathfrak{u}\}}_{h(\langle s,c\rangle)}, w^t_\mathfrak{u} \right\rangle$.

Consider next sentences of the form $\Diamond q$, for atomic q. In the update semantics, again, $\Diamond q$ is treated as a "test": it takes a context c and returns c unchanged just in case $[\![q]\!]_\mathfrak{u}(c) \neq \varnothing$, and otherwise returns $\varnothing$. That means that, for atomic q, $\langle s, c\rangle \in [\![\Diamond q]\!]_\mathfrak{u}$ iff

[48] For atomic p, a p-world is a world w where $v_\mathfrak{u}(p, w) = 1$; a $\bar{p}$-world is a world where $v_\mathfrak{u}(p, w) = 0$.

- (i) there is a q-world in s and $s = c$; then $\Diamond q$ is $\text{true}_\mathfrak{r}$ at $g(\langle s, c\rangle) = \left\langle f^s_{h(\langle s,c\rangle)*}, \{p : \forall w' \in s : p \in w'\}\right\rangle$, since $f^s_{h(\langle s,c\rangle)*}(w)$ will contain a q-world for any $w \neq w^t_\mathfrak{u}$; and $\{p : \forall w' \in s : p \in w'\} = w^t_\mathfrak{u}$ iff $s = \{w^t_\mathfrak{u}\}$, in which case by construction of h we have that $h(\langle s, c\rangle)$ contains $w^t_\mathfrak{u}$, and thus $f^s_{h(\langle s,c\rangle)*}(\{p : \forall w' \in s : p \in w'\})$ will then also contain a q-world, namely $w^t_\mathfrak{u}$;
- or (ii) there is no q-world in s, and $c = \varnothing$.
  - Suppose first that $s \neq c$. Then $\Diamond q$ will be $\text{true}_\mathfrak{r}$ at $g(\langle s, c\rangle) = \langle f^{\{w:\forall p:p\in w\rightarrow\forall w'\in s:p\notin w'\}}_{h(\langle s,c\rangle)*}$, $\{p : \forall w' \in s : p \notin w'\}\rangle$, since the fact that q is false throughout s ensures that q will be true at some world in $\{w : \forall p : p \in w \rightarrow \forall w' \in s : p \notin w'\}$, and $f^{\{w:\forall p:p\in w\rightarrow\forall w'\in s:p\notin w'\}}_{h(\langle s,c\rangle)*}$ takes every world but $w^t_\mathfrak{u}$ to $\{w : \forall p : p \in w \rightarrow \forall w' \in s : p \notin w'\}$. Moreover in this case we have $\{p : \forall w' \in s : p \notin w'\} = w^t_\mathfrak{u}$ iff $\langle s, c\rangle = \left\langle \{w^f_\mathfrak{u}\}, \varnothing\right\rangle$; then $h(\langle s, c\rangle) = \{w^t_\mathfrak{u}\}$ by construction of h, and so $f^{\{w:\forall p:p\in w\rightarrow\forall w'\in s:p\notin w'\}}_{h(\langle s,c\rangle)*}(\{p : \forall w' \in s : p \notin w'\})$ again contains a q-world.
  - Suppose next that $s = c$. Then $\Diamond q$ will be $\text{true}_\mathfrak{r}$ at $g(\langle s, c\rangle) = \left\langle f^{\{w^t_\mathfrak{u}\}}_{h(\langle s,c\rangle)}, w^t_\mathfrak{u}\right\rangle$, since $f^{\{w^t_\mathfrak{u}\}}_{h(\langle s,c\rangle)}(w^t_\mathfrak{u}) = \{w^t_\mathfrak{u}\}$.

Next, suppose that $\Diamond q$ is $\text{false}_\mathfrak{u}$ at $\langle s, c\rangle$; this will hold iff:

- s doesn't contain a q-world and $c \neq \varnothing$; then either
  - $s = c$; then $g(\langle s, c\rangle) = \left\langle f^s_{h(\langle s,c\rangle)*}, \{p : \forall w' \in s : p \in w'\}\right\rangle$. We know $\{p : \forall w' \in s : p \in w'\} \neq w^t_\mathfrak{u}$, for we would only have $\{p : \forall w' \in s : p \in w'\} = w^t_\mathfrak{u}$ if $\langle s, c\rangle = \langle\{w^t_\mathfrak{u}\}, \{w^t_\mathfrak{u}\}\rangle$, in which case s contains a q-world contrary to assumption; and so we have $f^s_{h(\langle s,c\rangle)*}(\{p : \forall w' \in s : p \in w'\}) = s$; since s doesn't contain a q-world, $\Diamond q$ is $\text{false}_\mathfrak{r}$ here;
  - or $c \subset s \wedge c \neq \varnothing$; then $g(\langle s, c\rangle) = \langle f^{\varnothing}_{h(\langle s,c\rangle)*}, \{p : (\forall w' \in c : p \in w') \wedge (\forall w'' \in (s \backslash c) : p \notin w'')\}\rangle$. Suppose first $\{p : (\forall w' \in c : p \in w') \wedge (\forall w'' \in (s \setminus c) : p \notin w'')\} \neq w^t_\mathfrak{u}$; so we have $f^{\varnothing}_{h(\langle s,c\rangle)*}(\{p : (\forall w' \in c : p \in w') \wedge (\forall w'' \in (s \setminus c) : p \notin w'')\}) = \varnothing$ and so $\Diamond q$ is $\text{false}_\mathfrak{r}$. Suppose next $\{p : (\forall w' \in c : p \in w') \wedge (\forall w'' \in (s \setminus c) : p \notin w'')\} = w^t_\mathfrak{u}$; then $\langle s, c\rangle$ must be $\left\langle \{w^t_\mathfrak{u}, w^f_\mathfrak{u}\}, \{w^t_\mathfrak{u}\}\right\rangle$; but then it contains a q-world, contrary to assumption;
  - or $c \not\subseteq s$; then $g(\langle s, c\rangle) = \left\langle f^{\varnothing}_{h(\langle s,c\rangle)*}, w^f_\mathfrak{u}\right\rangle$, and so $f^{\varnothing}_{h(\langle s,c\rangle)*}(w^f_\mathfrak{u}) = \varnothing$ and so $\Diamond q$ will be $\text{false}_\mathfrak{r}$ here;
- or s contains a q-world and $s \neq c$; then
  - either $c \subset s \wedge c \neq \varnothing$; then $g(\langle s, c\rangle) = \langle f^{\varnothing}_{h(\langle s,c\rangle)*}, \{p : (\forall w' \in c : p \in w') \wedge (\forall w'' \in (s \setminus c) : p \notin w'')\}\rangle$. Suppose first $\{p : (\forall w' \in c : p \in w') \wedge (\forall w'' \in (s \setminus c) : p \notin w'')\} \neq w^t_\mathfrak{u}$; so we have $f^{\varnothing}_{h(\langle s,c\rangle)*}(\{p : (\forall w' \in c : p \in w') \wedge (\forall w'' \in (s \setminus c) : p \notin w'')\}) = \varnothing$ and so $\Diamond q$ is $\text{false}_\mathfrak{r}$ here; suppose next $\{p : (\forall w' \in c : p \in w') \wedge (\forall w'' \in (s \setminus c) : p \notin w'')\} = w^t_\mathfrak{u}$, then $\langle s, c\rangle$ must be $\left\langle \{w^t_\mathfrak{u}, w^f_\mathfrak{u}\}, \{w^t_\mathfrak{u}\}\right\rangle$; by construction $h(\left\langle \{w^f_\mathfrak{u}, w^t_\mathfrak{u}\}, \{w^t_\mathfrak{u}\}\right\rangle) = \{w^f_\mathfrak{u}\}$; and so in this case, $f^{\varnothing}_{h(\langle s,c\rangle)*}(\{p : (\forall w' \in c : p \in w') \wedge (\forall w'' \in (s \setminus c) : p \notin w'')\}) = \{w^f_\mathfrak{u}\}$ and so again $\Diamond q$ is $\text{false}_\mathfrak{r}$;
  - or $c \not\subseteq s$; then $g(\langle s, c\rangle) = \left\langle f^{\varnothing}_{h(\langle s,c\rangle)*}, w^f_\mathfrak{u}\right\rangle$, and so $f^{\varnothing}_{h(\langle s,c\rangle)*}(w^f_\mathfrak{u}) = \varnothing$ and so $\Diamond q$ will be $\text{false}_\mathfrak{r}$ here;
  - or $s \neq c \wedge c = \varnothing$; then $g(\langle s, c\rangle) = \langle f^{\{w:\forall p:p\in w\rightarrow\forall w'\in s:p\notin w'\}}_{h(\langle s,c\rangle)*}, \{p : \forall w' \in s : p \notin w'\}\rangle$. Provided $\{p : \forall w' \in s : p \notin w'\} \neq w^t_\mathfrak{u}$, we have $f^{\{w:\forall p:p\in w\rightarrow\forall w'\in s:p\notin w'\}}_{h(\langle s,c\rangle)*}(\{p : \forall w' \in s : p \notin w'\}) = \{w :$

$\forall p : p \in w \rightarrow \forall w' \in s : p \notin w'\}$ which will not contain a q-world, since there is a q-world in s. And in this case we have $\{p : \forall w' \in s : p \notin w'\} = w_{\mathfrak{u}}^t$ iff $\langle s, c\rangle = \left\langle \{w_{\mathfrak{u}}^f\}, \varnothing\} \right\rangle$, contrary to our assumption that s contains a q-world.

Finally, in any update model, for any $\varphi \in \mathcal{L}^{\Diamond}$ and any update index i, $\Diamond(\Diamond\varphi)$ is true$_{\mathfrak{u}}$ at i iff $\Diamond\varphi$ is true$_{\mathfrak{u}}$ at i. And this will also hold relative to any point in the image of g:

- When $s = c = \varnothing$, then both $\Diamond\varphi$ and $\Diamond(\Diamond\varphi)$ will be true$_{\mathfrak{r}}$ (by an obvious induction on the length of formulas which I leave implicit here);
- when $c \not\subseteq s$, both will be false$_{\mathfrak{r}}$;
- when $c \subset s \wedge c \neq \varnothing$, $g_1(\langle s, c\rangle)(g_2(\langle s, c\rangle))$ is empty, making both false; unless $g_2(\langle s, c\rangle) = w_{\mathfrak{u}}^t$, which holds only when $s = \{w_{\mathfrak{u}}^f, w_{\mathfrak{u}}^t\}$ and $c = \{w_{\mathfrak{u}}^t\}$, in which case by construction of h, we have $g_1(\langle s, c\rangle)(g_2(\langle s, c\rangle)) = \{w_{\mathfrak{u}}^f\}$.[49] Since $g_1(\langle s, c\rangle)(w_{\mathfrak{u}}^f) = \varnothing$, we again have false;
- when $s = c \neq \varnothing$ and $g_2(\langle s, c\rangle) \neq w_{\mathfrak{u}}^t$, we have $g_1(\langle s, c\rangle)$ has the same value at $g_2(\langle s, c\rangle)$ and at every element of $g_1(\langle s, c\rangle)(g_2(\langle s, c\rangle))$, provided s does not include $w_{\mathfrak{u}}^t$; and whenever s includes $w_{\mathfrak{u}}^t$, we have $g_1(\langle s, c\rangle)(w_{\mathfrak{u}}^t)$ includes $w_{\mathfrak{u}}^t$ by construction of h, and thus that $\forall w \in s : w_{\mathfrak{u}}^t \in g_1(\langle s, c\rangle)(w)$, and thus that $\Diamond\varphi$ is true at $g_2(\langle s, c\rangle)$ and at every world in $g_1(\langle s, c\rangle)(g_2(\langle s, c\rangle))$; and $g_2(\langle s, c\rangle) = w_{\mathfrak{u}}^t$ iff $s = c = \{w_{\mathfrak{u}}^t\}$, in which case $\Diamond\varphi$ is true for any φ at $g(\langle s, c\rangle)$;
- when $s \neq c \wedge c = \varnothing$, $g_2(\langle s, c\rangle) = w_{\mathfrak{u}}^t$ iff $s = \{w_{\mathfrak{u}}^f\}$; then $h(\langle s, c\rangle) = \{w_{\mathfrak{u}}^t\}$, so then $\Diamond\varphi$ is true at $g(\langle s, c\rangle)$ for all φ. When $g_2(\langle s, c\rangle) \neq w_{\mathfrak{u}}^t$, we know that $w_{\mathfrak{u}}^t \notin \{w : \forall p : p \in w \rightarrow \forall w' \in s : p \notin w'\}$ and so $g_1(\langle s, c\rangle)(g_2(\langle s, c\rangle))$ is the same as $g_1(\langle s, c\rangle)$ applied to any element of $g_1(\langle s, c\rangle)(g_2(\langle s, c\rangle))$.

And thus we can conclude that $\Diamond\varphi$ is true$_{\mathfrak{u}}$ at i iff $\Diamond\varphi$ is true$_{\mathfrak{r}}$ at $g(i)$, for any $\varphi \in \mathcal{L}^{\Diamond}$. Thus g is an injection from the indices of $\mathfrak{u}$ to those of $\mathfrak{r}$ which preserves truth for all sentences of $\mathcal{L}^{\Diamond}$, and thus by Fact 3.1 we have $\mathcal{U} \preceq_{\mathcal{L}^{\Diamond}} \mathcal{R}$.[50]

The proof that $\mathcal{R} \not\preceq_{\mathcal{L}^{\Diamond}} \mathcal{U}$ will be as for Fact 3.3.

□

Fact 3.9. $\mathcal{U} \not\preceq_{\mathcal{L}^{\Diamond}} \mathcal{D}$.

Proof. Consider arbitrary $\mathfrak{u} \in \mathcal{U}$ and $\mathfrak{d} \in \mathcal{D}$. Consider the three $\mathfrak{u}$-indices $\langle \varnothing, \{w\}\rangle, \langle \varnothing, \{w'\}\rangle, \langle \varnothing, \{w''\}\rangle$, with w, w', and w'' all different. These three indices all make p false$_{\mathfrak{u}}$, for any atomic $p \in \mathcal{L}^{\Diamond}$; they also make $\Diamond\varphi$ false$_{\mathfrak{u}}$, for any $\varphi \in \mathcal{L}^{\Diamond}$, and thus make every sentence in $\mathcal{L}^{\Diamond}$ false$_{\mathfrak{u}}$. There are, however, only two indices in $\mathfrak{d}$ which make every sentence in $\mathcal{L}^{\Diamond}$ false$_{\mathfrak{d}}$, namely $\left\langle \varnothing, w_{\mathfrak{d}}^f \right\rangle$ and $\left\langle \{w_{\mathfrak{d}}^f\}, w_{\mathfrak{d}}^f \right\rangle$, where $w_{\mathfrak{d}}^f$ is the world where, for every atomic sentence p, $v_{\mathfrak{d}}(p, w_{\mathfrak{d}}^f) = 0$. Thus any truth-preserving function $g : \mathfrak{I}_{\mathfrak{u}} \rightarrow \mathfrak{I}_{\mathfrak{d}}$ will have to map at least two of the $\mathfrak{u}$-indices in question to the same $\mathfrak{d}$-index, and thus will fail to be an injection. Thus Fact 3.9 follows by Fact 3.1.

□

Fact 3.10. $\mathcal{D} \not\preceq_{\mathcal{L}^{\Diamond}} \mathcal{U}$.

[49] g_n is the n^{th} projection of g, i.e. $g_n(X) = x_n$ iff $g(X) = \langle x_1, x_2, \ldots x_n, \ldots\rangle$.

[50] Note that not every operator which can be added to $\mathcal{U}$ will be well-defined if we want the intension of any sentence in $\mathcal{U}$ to be a function from contexts to contexts, rather than just a relation; there are different approaches within broadly update-style frameworks to this question (e.g. Heim 1983 vs. Groenendijk and Stokhof 1991).

Proof. Consider arbitrary $\mathfrak{u} \in \mathcal{U}$ and $\mathfrak{d} \in \mathcal{D}$. There are exactly three $\mathfrak{u}$-indices which make all sentences in $\mathcal{L}^{\Diamond}$ true$_{\mathfrak{u}}$, namely $\langle \varnothing, \varnothing \rangle$, $\left\langle \{w_{\mathfrak{u}}^f\}, \varnothing \right\rangle$, and $\langle \{w_{\mathfrak{u}}^t\}, \{w_{\mathfrak{u}}^t\} \rangle$, where $w_{\mathfrak{u}}^f$ is the world where, for every atomic sentence p, $v_{\mathfrak{u}}(p, w_{\mathfrak{u}}^f) = 0$ and where $w_{\mathfrak{u}}^t$ is the world where, for every atomic sentence p, $v_{\mathfrak{u}}(p, w_{\mathfrak{u}}^t) = 1$. There are more than three $\mathfrak{d}$-indices which make all sentences in $\mathcal{L}^{\Diamond}$ true$_{\mathfrak{d}}$: these include $\langle \mathcal{W}_{\mathfrak{d}}, w_{\mathfrak{d}}^t \rangle$ and $\langle \{w_{\mathfrak{d}}^t\}, w_{\mathfrak{d}}^t \rangle$, as well as $\langle s, w_{\mathfrak{d}}^t \rangle$ for any s such that $\{w_{\mathfrak{d}}^t\} \subseteq s \subseteq \mathcal{W}_{\mathfrak{d}}$, with $w_{\mathfrak{d}}^t$ defined as for $w_{\mathfrak{u}}^t$. Thus any function from the indices of $\mathfrak{d}$ to those of $\mathfrak{u}$ which preserves truth for all sentences in $\mathcal{L}^{\Diamond}$ will have to map more than three $\mathfrak{d}$-indices to three $\mathfrak{u}$-indices, and so will fail to be an injection. Since $\mathfrak{d}$ and $\mathfrak{u}$ were chosen arbitrarily, by Lemma A.2 and Fact 3.1, $\mathcal{D} \not\preceq_{\mathcal{L}^{\Diamond}} \mathcal{U}$.

□

Fact 3.12. $\mathcal{S} \prec_{\mathcal{L}^{\Diamond}} \mathcal{D}$.

Proof. Recall that $\mathcal{S}$ is the set of models $\mathfrak{s}$ of the form $\langle \mathcal{W}_{\mathfrak{s}}, v_{\mathfrak{s}}, \mathfrak{I}_{\mathcal{S}}, [\![\cdot]\!]_{\mathfrak{s}} \rangle$ with $\mathfrak{I}_{\mathfrak{s}} = \{s : s \subseteq \mathcal{W}_{\mathfrak{s}}\}$, with $[\![\cdot]\!]_{\mathfrak{s}}$ specified as in §3, and assuming again that in any model, any two worlds differ on the truth of some atom, and for any set of atoms, exactly that set is true at some world. Note again that $\mathcal{S}$ is isomorphic with respect to $\mathcal{L}^{\Diamond}$. Chose a model $\mathfrak{s} \in \mathcal{S}$ and find a model $\mathfrak{d} \in \mathcal{D}$ s.t. $\mathcal{W}_{\mathfrak{s}} = \mathcal{W}_{\mathfrak{d}}$ and $v_{\mathfrak{s}} = v_{\mathfrak{d}}$. For convenience we identify every world in $\mathcal{W}_{\mathfrak{s}}$ with the set of atomic sentences it makes true according to $v_{\mathfrak{s}}$. Let the function g take any information state $s \subseteq \mathcal{W}_{\mathfrak{s}}$ to $\langle s, \bigcap s \rangle$. For atomic p, p is true$_{\mathfrak{s}}$ at s iff for all $w' \in s$, $v_{\mathfrak{s}}(p, w') = 1$, iff $p \in \bigcap s$, iff p is true$_{\mathfrak{d}}$ at $g(s) = \langle s, \bigcap s \rangle$. For atomic p, $\Diamond p$ is true$_{\mathfrak{s}}$ at s iff s contains a p-world (according to $v_{\mathfrak{s}}$) iff $\Diamond p$ is true$_{\mathfrak{d}}$ at $g(s) = \langle s, \bigcap s \rangle$. Finally, it holds in both $\mathfrak{s}$ and $\mathfrak{d}$ that $\Diamond(\Diamond\varphi)$ is true at an index iff $\Diamond\varphi$ is, and so we know that for any $\varphi \in \mathcal{L}^{\Diamond}$, $\Diamond\varphi$ will be true$_{\mathfrak{s}}$ at i iff $\Diamond\varphi$ is true$_{\mathfrak{d}}$ at $g(i)$. Note finally that g is an injection: for any s and s', if $s \neq s'$, then the first elements of $g(s)$ and $g(s')$ will differ. Thus by Lemma A.2 we have $\mathfrak{s} \preceq_{\mathcal{L}^{\Diamond}} \mathfrak{d}$ and so by Fact 3.1 we have $\mathcal{S} \preceq_{\mathcal{L}^{\Diamond}} \mathcal{D}$.

But we do not have the converse: $\mathcal{D} \not\preceq_{\mathcal{L}^{\Diamond}} \mathcal{S}$. Consider any models $\mathfrak{s} \in \mathcal{S}$ and $\mathfrak{d} \in \mathcal{D}$. Consider three $\mathfrak{d}$-indices $\langle \varnothing, w \rangle$ and $\langle \varnothing, w' \rangle$, and $\langle \varnothing, w'' \rangle$ with w, w', and w'' all distinct. For any $\varphi \in \mathcal{L}^{\Diamond}$, $\Diamond\varphi$ is false$_{\mathfrak{d}}$ at all these indices. The only indices which make $\Diamond\varphi$ false$_{\mathfrak{s}}$ for every $\varphi \in \mathcal{L}^{\Diamond}$ are $\varnothing$ and $\{w_{\mathfrak{s}}^f\}$, where $w_{\mathfrak{s}}^f$ is the world which makes every atom false according to $v_{\mathfrak{s}}$, and thus any truth-preserving function will have to take two of the $\mathfrak{d}$-indices to the same $\mathfrak{s}$ index, and thus will fail to be an injection; thus by Lemma A.2 and Fact 3.1, $\mathcal{D} \not\preceq_{\mathcal{L}^{\Diamond}} \mathcal{S}$.

□

Fact 3.13. For any language $\mathcal{L}$ and a set of semantic theories $\mathfrak{T}$ each of which is isomorphic with respect to $\mathcal{L}$, $\preceq_{\mathcal{L}}$ is a partial pre-order over $\mathfrak{T}$.

Proof. $\preceq_{\mathcal{L}}$ will be transitive: suppose $\mathcal{T} \preceq_{\mathcal{L}} \mathcal{T}'$, witnessed by a function $g : \mathfrak{I}_{\mathfrak{t}} \to \mathfrak{I}_{\mathfrak{t}'}$, for $\mathfrak{t} \in \mathcal{T}, \mathfrak{t}' \in \mathcal{T}'$, and $\mathcal{T}' \preceq_{\mathcal{L}} \mathcal{T}''$, witnessed by a function $f : \mathfrak{I}_{\mathfrak{t}*'} \to \mathfrak{I}_{\mathfrak{t}''}$, for $\mathfrak{t}*' \in \mathcal{T}', \mathfrak{t}'' \in \mathcal{T}''$. By isomorphism, there is a truth-preserving injection $h : \mathfrak{I}_{\mathfrak{t}'} \to \mathfrak{I}_{\mathfrak{t}*'}$. $f \circ (h \circ g)$ will witness $\mathcal{T} \preceq_{\mathcal{L}} \mathcal{T}''$. $\preceq_{\mathcal{L}}$ is reflexive, witnessed by the identity function. It is not anti-symmetric, since it is easy to see that there are different semantic theories $\mathcal{T}$ and $\mathcal{T}'$ isomorphic with respect to $\mathcal{L}$ such that $\mathcal{T} \preceq_{\mathcal{L}} \mathcal{T}'$ and $\mathcal{T}' \preceq_{\mathcal{L}} \mathcal{T}$ (for instance, two standard semantic theories for a language just comprising atomic sentences, with the same set of possible worlds but different valuations, may have this property). And it is not necessarily connected, since, as we saw in the comparison of $\mathcal{D}$ to $\mathcal{U}$, there are semantic theories which are incommensurable with respect to a given language.

□

Fact 3.15. $\mathcal{D}^{\exists} \prec_{\mathcal{L}_v^{\Diamond}} \mathcal{R}^{\exists}$.

Proof. Note that $\mathcal{D}^{\exists}$ and $\mathcal{R}^{\exists}$ are both isomorphic with respect to $\mathcal{L}_v^{\Diamond}$. Consider a model $\mathfrak{d}^{\exists} \in \mathcal{D}^{\exists}$. Find a model $\mathfrak{r}^{\exists} \in \mathcal{R}^{\exists}$ built on the same domain, set of worlds, and valuation function. For any index $\langle a, s, w \rangle \in \mathfrak{I}_{\mathfrak{d}^{\exists}}$, with a a variable assignment, s a set of worlds, and w a world, let g be the function which takes $\langle a, s, w \rangle$ to the $\mathfrak{r}^{\exists}$-index

$\langle a, f^s, w\rangle$, where f^s is again the constant function to s. g will witness $\mathfrak{d}^\exists \preceq_{\mathcal{L}_v^\Diamond} \mathfrak{r}^\exists$; the proof is a generalization of the parallel result in the proof of Fact 3.2; and so we have $\mathcal{D}^\exists \preceq_{\mathcal{L}_v^\Diamond} \mathcal{R}^\exists$. But there is no truth-preserving injection in the other direction, for the same reasons given in the proof of Fact 3.3. Note moreover that the proof of Fact 3.5 can be extended to show that $\mathcal{D}^\exists \prec_{\mathcal{L}_v^{\Diamond-}} \mathcal{R}^\exists$.

□

Fact 3.16. $\mathcal{U}^\exists \prec_{\mathcal{L}_v^\Diamond} \mathcal{R}^\exists$.

Proof. First note $\mathcal{U}^\exists$ is isomorphic with respect to $\mathcal{L}_v^\Diamond$. Then our proof is very much as in the proof of Fact 3.6. We construct a witness function g from an arbitrarily chosen model $\mathfrak{u}^\exists \in \mathcal{U}^\exists$ to a model $\mathfrak{r}^\exists \in \mathcal{R}^\exists$ built on the same set of worlds, valuation function, and domain, using the function h defined in the proof of Fact 3.6; with the notation defined there:

$$g(\langle a, \langle s, c\rangle\rangle) = \begin{cases} \left\langle a, f^s_{h(\langle s,c\rangle)*}, \iota w : \forall R : v_{\mathfrak{u}^\exists}(R, w) = \bigcap\{v_{\mathfrak{u}^\exists}(R, w') : w' \in s\}\right\rangle & \text{iff } s = c \neq \varnothing \\ \langle a, f^{\varnothing}_{h(\langle s,c\rangle)*}, \iota w : \forall R : v_{\mathfrak{u}^\exists}(R, w) = & \\ \quad \bigcap\{v_{\mathfrak{u}^\exists}(R, w') : w' \in c\} \setminus \bigcup\{v_{\mathfrak{u}^\exists}(R, w') : w' \in (s \setminus c)\}\rangle & \text{iff } c \subset s \wedge c \neq \varnothing \\ \left\langle a, f^{\varnothing}_{h(\langle s,c\rangle)*}, w^f_{\mathfrak{u}^\exists}\right\rangle & \text{iff } c \not\subseteq s \\ \langle a, f^{\{w:\forall R\forall \vec{d}:\vec{d}\in v_{\mathfrak{u}^\exists}(R,w)\rightarrow\forall w'\in s:\vec{d}\notin v_{\mathfrak{u}^\exists}(R,w')\}}_{h(\langle s,c\rangle)*}, & \\ \quad \iota w : \forall n : \forall R^n : v_{\mathfrak{u}^\exists}(R^n, w) = D^n \setminus \bigcup\{v_{\mathfrak{u}^\exists}(R^n, w') : w' \in s\}\rangle & \text{iff } s \neq c \wedge c = \varnothing \\ \left\langle a, f^{\{w^t_{\mathfrak{u}^\exists}\}}_{h(\langle s,c\rangle)}, w^t_{\mathfrak{u}^\exists}\right\rangle & \text{iff } s = c = \varnothing \end{cases}$$

R ranges over relation symbols in the vocabulary of $\mathcal{L}_v^\Diamond$; $w^t_{\mathfrak{u}^\exists}$ is the world such that $v_{\mathfrak{u}^\exists}(R^n, w^t_{\mathfrak{u}^\exists})$ is the universal n-ary relation, for any n-place relation symbol R^n; and $w^f_{\mathfrak{u}^\exists}$ the world such that $v_{\mathfrak{u}^\exists}(R^n, w^f_{\mathfrak{u}^\exists})$ is the empty relation, for any R^n. D is the domain of individuals; $\vec{d}$ ranges over ordered sequences of elements of D; and ι ranges over worlds in $\mathcal{W}_{\mathfrak{u}^\exists}$. The proof that g is a witness function is parallel to the proof of Fact 3.6. The proof that $\mathcal{R}^\exists \not\preceq_{\mathcal{L}_v^\Diamond} \mathcal{U}^\exists$ is parallel to that for Fact 3.3.

□

References

Aloni, M. (2000). Conceptual covers in dynamic semantics. In Cavedon, L., Blackburn, P., Braisby, N., and Shimojima, A., editors, *Logic, Language and Computation*, volume III.

Aloni, M. (2016). FC disjunction in state-based semantics. Slides for Logical Aspects of Computational Linguistics (*LACL*), Nancy, France.

Aloni, M. D. (2001). *Quantification Under Conceptual Covers*. PhD thesis, University of Amsterdam, Amsterdam.

Beaver, D. (1994). When variables don't vary enough. In Harvey, M. and Santelmann, L., editors, *Semantics and Linguistic Theory (SALT)*, volume 4, pages 35–60.

Beaver, D. (2001). *Presupposition and Assertion in Dynamic Semantics*. CSLI Publications: Stanford, CA.

Beaver, D. I. (1992). The kinematics of presupposition. In *ITLI Prepublication Series for Logic, Semantics and Philosophy of Language*. University of Amsterdam.

Beddor, B. and Goldstein, S. (2018). Believing epistemic contradictions. *Review of Symbolic Logic*, 11(1):87–114.

Bledin, J. and Lando, T. (2017). Closure and epistemic modals. To appear in *Philosophy and Phenomenological Research*.

Cresswell, M. (1990). *Entities and Indices*. Kluwer Academic Publishers, Dordrecht.
Degen, J., Kao, J. T., Scontras, G., and Goodman, N. D. (2015). A cost- and information-based account of epistemic *must*. Poster at 28th Annual CUNY Conference on Human Sentence Processing.
Dorr, C. and Hawthorne, J. (2013). Embedding epistemic modals. *Mind*, 122(488):867–913.
Dowell, J. (2011). A flexibly contextualist account of epistemic modals. *Philosophers' Imprint*, 11(14):1–25.
Egan, A., Hawthorne, J., and Weatherson, B. (2005). Epistemic modals in context. In Preyer, G. and Peter, G., editors, *Contextualism in Philosophy: Knowledge, Meaning and Truth*, chapter 6, pages 131–169. Oxford University Press.
von Fintel, K. (1997). Bare plurals, bare conditionals, and *Only*. *Journal of Semantics*, 14:1–56.
von Fintel, K. and Gillies, A. (2010). Must...stay...strong! *Natural Language Semantics*, 18(4):351–383.
French, R. (2017). Notational variance and its variants. *Topoi*, pages 1–11.
Gerbrandy, J. (1998). Identity in epistemic semantics. In *Third Conference on Information Theoretic Approaches to Logic, Language and Computation*.
Giannakidou, A. and Mari, A. (2016). Epistemic future and epistemic MUST: nonveridicality, evidence, and partial knowledge. In Blaszack, J., Giannikidou, A., Klimek-Jankowska, D., and Mygdalski, K., editors, *Mood, Aspect and Modality: What is a Linguistic Category?* University of Chicago Press.
Gillies, A. S. (2018). Updating data semantics. *Mind*.
Groenendijk, J. and Stokhof, M. (1991). Dynamic predicate logic. *Linguistics and Philosophy*, 14(1):39–100.
Groenendijk, J., Stokhof, M., and Veltman, F. (1996). Coreference and modality. In *Handbook of Contemporary Semantic Theory*, pages 179–216. Oxford: Blackwell.
Hacquard, V. (2006). *Aspects of Modality*. PhD thesis, MIT.
Hacquard, V. (2010). On the event relativity of modal auxiliaries. *Natural Language Semantics*, 18:79–114.
Hawke, P. and Steinert-Threlkeld, S. (2016). Informational dynamics of epistemic possibility modals. *Synthese*.
Hawthorne, J., Rothschild, D., and Spectre, L. (2016). Belief is weak. *Philosophical Studies*, 173(5):1393–1404.
Heim, I. (1982). *The Semantics of Definite and Indefinite Noun Phrases*. PhD thesis, University of Massachusetts, Amherst.
Heim, I. (1983). On the projection problem for presuppositions. In Barlow, M., Flickinger, D. P., and Wiegand, N., editors, *The West Coast Conference on Formal Linguistics (WCCFL)*, volume 2, pages 114–125. Stanford, Stanford University Press.
Heim, I. (1992). Presupposition projection and the semantics of attitude verbs. *Journal of Semantics*, 9(3):183–221.
Hintikka, J. (1962). *Knowledge and Belief: An Introduction to the Logic of Two Notions*. Cornell University Press, Ithaca, NY.
Holliday, W. H. and Icard, III, T. F. (2017). Indicative conditionals and dynamic epistemic logic. In *Theoretical Aspects of Rationality and Knowledge (TARK)*, volume 16.
Ippolito, M. (2017). Constraints on the embeddability of epistemic modals. In Truswell, R., Cummins, C., Heycock, C., Rabern, B., and Rohde, H., editors, *Sinn und Bedeutung*, volume 21, pages 605–622.
Kamp, H. (1970). Formal properties of 'now'. *Theoria*, 37(227-273).
Karttunen, L. (1972). Possible and must. In Kimball, J., editor, *Syntax and Semantics*, volume 1, pages 1–20. Academic Press, New York.
Khoo, J. (2015). Modal disagreements. *Inquiry*, 58(5):511–534.
Kratzer, A. (1977). What 'must' and 'can' must and can mean. *Linguistics and Philosophy*, 1(3):337–355.
Kratzer, A. (1981). The notional category of modality. In Eikmeyer, H. and Rieser, H., editors, *Words, Worlds, and Contexts: New Approaches in Word Semantics*, pages 38–74. de Gruyter.
Kratzer, A. (1991). Modality. In von Stechow, A. and Wunderlich, D., editors, *Semantics: An International Handbook of Contemporary Research*, pages 639–650. de Gruyter, Berlin.
Kratzer, A. (2012). *Modals and Conditionals*. Oxford University Press, Oxford.
Kripke, S. (1963). Semantical considerations on modal logic. *Acta Philosophica Fennica*, 16:83–94.
Lassiter, D. (2016). *Must*, knowledge, and (in)directness. *Natural Language Semantics*.
Lewis, D. (1980). Index, context, and content. In Kanger, S. and Ohman, S., editors, *Philosophy and Grammar*, pages 79–100. D. Reidel.
MacFarlane, J. (2011). Epistemic modals are assessment sensitive. In Egan, A. and Weatherson, B., editors,

Epistemic Modality. Oxford University Press.
MacFarlane, J. (2014). *Assessment Sensitivity: Relative Truth and Its Applications*. Oxford University Press, Oxford.
Mandelkern, M. (2017a). *Coordination in Conversation*. PhD thesis, Massachusetts Institute of Technology.
Mandelkern, M. (2017b). A solution to Karttunen's problem. In Truswell, R., Cummins, C., Heycock, C., Rabern, B., and Rohde, H., editors, *Sinn und Bedeutung 21*, pages 827–844.
Mandelkern, M. (2018a). How to do things with modals. To appear in *Mind & Language*.
Mandelkern, M. (2018b). What 'must' adds. To appear in *Linguistics and Philosophy*.
Mandelkern, M. (2019). Bounded modality. *The Philosophical Review*, 181(1).
Matthewson, L. (2015). Evidential restrictions on epistemic modals. In Alonso-Ovalle, L. and Menendez-Benito, P., editors, *Epistemic Indefinites*. Oxford University Press, New York.
Moss, S. (2015). On the semantics and pragmatics of epistemic vocabulary. *Semantics and Pragmatics*, 8(5):1–81.
Mossakowski, T., Diaconescu, R., and Tarlecki, A. (2009). What is a logic translation? *Logica Universalis*, 3(1):95–124.
Ninan, D. (2010). Semantics and the objects of assertion. *Linguistics and Philosophy*, 33(5):355–380.
Ninan, D. (2016). Relational semantics and domain semantics for epistemic modals. *Journal of Philosophical Logic*, 47(1):1–16.
Ninan, D. (2018). Quantification and epistemic modality. *The Philosophical Review*, 127(2):433–485.
Peters, S. and Westerståhl, D. (2008). *Quantifiers in Language and Logic*. Oxford University Press.
Pinheiro Fernandes, D. (2017). Translations: generalizing relative expressiveness between logics. Manuscript, University of Salamanca. `https://arxiv.org/abs/1706.08481`.
Rabern, B. (2012). Against the identification of assertoric content with compositional value. *Synthese*, 189(1):75–96.
Rabern, B. (2013). Monsters in Kaplan's logic of demonstratives. *Philosophical Studies*, 164:393–404.
Rothschild, D. (2011). Expressing credences. In *Proceedings of the Aristotelian Society*, volume 112, pages 99–114.
Rothschild, D. (2017). Veltman-Yalcin. http://danielrothschild.com/dyncon/vy/.
Rothschild, D. and Klinedinst, N. (2015). Quantified epistemic modality. Handout for talk at Birmingham.
Rothschild, D. and Yalcin, S. (2015). On the dynamics of conversation. *Noûs*, 51(1):24–48.
Rothschild, D. and Yalcin, S. (2016). Three notions of dynamicness in language. *Linguistics and Philosophy*, 39(4):333–355.
Schultz, M. (2010). Epistemic modals and informational consequence. *Synthese*, 174(3):385–395.
Sherman, B. (2018). Open questions and epistemic necessity. *The Philosophical Quarterly*.
Stalnaker, R. (1984). *Inquiry*. MIT.
Steinert-Threlkeld, S. (2017). *Communication and Computation: New Questions About Compositionality*. PhD thesis, Stanford University.
Stephenson, T. (2007). Judge dependence, epistemic modals, and predicates of personal taste. *Linguistics and Philosophy*, 30(4):487–525.
Swanson, E. (2015). The application of constraint semantics to the language of subjective uncertainty. *Journal of Philosophical Logic*, 45(121):121–146.
Veltman, F. (1985). *Logics for Conditionals*. PhD thesis, University of Amsterdam.
Veltman, F. (1996). Defaults in update semantics. *Journal of Philosophical Logic*, 25(3):221–261.
Willer, M. (2013). Dynamics of epistemic modality. *Philosophical Review*, 122(1):45–92.
Yalcin, S. (2007). Epistemic modals. *Mind*, 116(464):983–1026.
Yalcin, S. (2011). Nonfactualism about epistemic modality. In Egan, A. and Weatherson, B., editors, *Epistemic Modality*, pages 295–332. Oxford University Press.
Yalcin, S. (2012). Context probabilism. In Aloni, M., Kimmelman, V., Roelofsen, F., Sassoon, G. W., Schulz, K., and Westera, M., editors, *The 18th Amsterdam Colloquium*, pages 12–21.
Yalcin, S. (2015). Epistemic modality *de re*. *Ergo*, 2(19):475–527.