Knowing What Follows:

Epistemic Closure and Epistemic Logic

Wesley H. Holliday

ILLC Dissertation Series DS-2012-09

Institute for Logic, Language and Computation

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

KNOWING WHAT FOLLOWS:
EPISTEMIC CLOSURE AND EPISTEMIC LOGIC

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF PHILOSOPHY
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Wesley H. Holliday
Original Stanford Version - June 2012
Revised ILLC Version - December 2013

Advisors:
Prof. Johan van Benthem
Prof. Krista Lawlor

Readers:
Prof. Helen Longino
Prof. Eric Pacuit

Defense:
May 22, 2012

Department of Philosophy
Stanford University
Stanford, CA
USA

Abstract

The starting point of this dissertation is a central question in epistemology and epistemic logic, statable roughly as follows: is it a necessary condition of an agent's knowing some propositions $P_1, \ldots, P_n$ that she has done enough empirical investigation of the world so that she could know *any logical consequence* of $\{P_1, \ldots, P_n\}$ without further empirical investigation? An affirmative answer amounts to a claim of full *epistemic closure*: the set of propositions that an agent knows or could know without further empirical investigation is closed under multi-premise logical consequence.

The idea of full epistemic closure creates a tension with an attractive idea of *fallibilism* about knowledge. According to fallibilism, for an agent to know a true empirical proposition P, it is not required that her evidence rules out *every possible way* in which P could be false and some incompatible alternative hypothesis could obtain. If such a feat were required, agents would know almost nothing. Yet full epistemic closure requires for knowledge of P that an agent does know—or could know without further empirical investigation—the negation of every such alternative hypothesis, assuming she knows that these hypotheses are incompatible with P. Although not a formal contradiction between closure and fallibilism, this is a tension to say the least.

In this dissertation, I explore the extent to which it is possible to make fallibilism compatible with closure. I begin by formalizing a family of fallibilist theories of knowledge in models for epistemic logic. Model-theoretic methods are used to characterize the closure properties of knowledge according to different fallibilist pictures, identify the structural features of these pictures that correspond to closure properties, transform models of one theory into models of another, prove impossibility results, and ultimately find a middle way between full closure and no closure for fallibilism.

I argue that the standard versions of "Fallibilism 1.0" each face one of three serious problems related to closure: the Problem of Vacuous Knowledge, the Problem of Containment, and the Problem of Knowledge Inflation. To solve these problems, I propose a new framework for Fallibilism 2.0: the Multipath Picture of Knowledge. This picture is based on taking seriously the idea that there can be multiple paths to knowing a complex claim about the world. An overlooked consequence of fallibilism is that these multiple paths to knowledge may involve ruling out different sets of alternatives, which should be represented in our picture of knowledge. I argue that the Multipath Picture of Knowledge is a better picture for all fallibilists, whether for or against full closure. Yet I also argue that only by accepting less than full closure can we solve the closure-related problems that plague previous versions of fallibilism.

Acknowledgements

It has been a privilege to be a graduate student at Stanford University for the past five years. Although I had already been at Stanford for four years as an undergraduate, my first year as a graduate student was another kind of reawakening. In the fall and winter, I took my first courses from Krista Lawlor and Dagfinn Føllesdal, from whom I have learned so much about epistemology and the philosophy of language. I will also never forget the intellectual excitement of taking the Stanford logic sequence from Marc Pauly, philosophy of mathematics from Solomon Feferman, and sitting in on Johan van Benthem's advanced modal logic seminar in the winter and spring.

Since then I have had more wonderful teachers in logic, including Eric Pacuit, Grigori Mints, and Fernando Ferreira, and I have had the pleasure of working with many enthusiastic students as a TA or instructor for eight logic courses. While the focus of my research has been on logic and epistemology, I have also been fortunate to have had oustanding teachers in the philosophy of science, including Helen Longino, Michael Friedman, Patrick Suppes, and Thomas Ryckman, and serving as Tom's TA was another pleasure. As both an undergraduate and a graduate student, I have learned a great deal about a variety of philosophical topics from Graciela De Pierris and Nadeem Hussain. I have also benefited from interacting with philosophers outside of Stanford, especially John Martin Fischer over the past five years and Sherri Roush in more recent months, not to forget my formative undergraduate interactions with George Tsai and Georges Rey and their encouragement about graduate school. I have also learned from logicians on other continents through correspondence and conferences, especially Hans van Ditmarsch, Alexandru Baltag, and Sonja Smets.

In all of my nine years at Stanford, John Perry has been a source of the deepest

philosophical lessons, humor, and encouragement. I cannot imagine having come this far without his wise advice about my work and education, for which I will always be grateful. I also owe tremendous thanks to my dissertation committee: my co-advisors, Johan van Benthem and Krista Lawlor, and my readers, Helen Longino and Eric Pacuit. My first serious encounter with the problems discussed in this dissertation was in Krista's superb M&E Core Seminar in Fall 2007, and my first serious work with the modal-logical techniques that I bring to bear on these problems came from Johan's masterful modal logic course in Spring 2009. Being a student of Johan, who radiates energy and a sense of possibilities and progress, has been an inspiration. The feedback I have received from Johan on my logical research has been invaluable, thanks to his uncanny ability to see connections and creative twists in so many places. Krista has also been an incredible advisor, and I cannot think of a better place to bounce ideas around than Krista's office. Her insight has been essential for my understanding of the epistemological issues I discuss in this dissertation, and her mentoring has been essential for preparing me for academic life after Stanford.

As if I were not lucky enough to have Johan and Krista as advisors, I also had the good fortune to have Helen and Eric on my committee. My path toward a dissertation in epistemology began with an idea for a dissertation in social epistemology, stimulated by reading groups with Helen and Eric. While the path turned toward individual epistemology, I was very glad to have had the guidance of Helen and Eric along the way. It has also been an honor to know Dagfinn Føllesdal, whose dissertation and subsequent work shows that logical investigation can have deep philosophical significance, a motivating idea behind this dissertation. In addition to interacting with these fantastic faculty, I have gained much from interacting with previous Stanford students in logic, epistemology, and related areas, including Jesse Alama, Alexei Angelides, Alistair Isaac, Neil Van Leeuwen, and Teru Miyake, and from current students including Peter Hawke and Tal Glezer. Finally, some of the best preparation for writing this dissertation has come from conversations and collaborations with Tomohiro Hoshi and Thomas Icard [Holliday and Icard, 2010, Holliday et al., 2011, 2012]. Working on open problems in our Logical Dynamics Lab at CSLI has been great fun, as was taking this research "on the road" with Thomas from Indiana to Copenhagen.

For funding I received for a series of extremely productive trips, I am very grateful to Helen Longino as Chair of Philosophy, to Johan van Benthem, and to a Graduate Research Opportunity Award from Stanford H&S. Some of the first ideas that lead to this dissertation occurred to me at the European Summer School for Logic, Language, and Information (ESSLLI) 2009 in Bordeaux. A few months later, I participated in the LORI-II Conference in Chongqing, China, which opened my eyes to the wide world of 21st century logic. After a week of fascinating classes at NASSLLI 2010 in Indiana—for which I gratefully acknowledge support from the University of Indiana—I presented the first paper related to what would become this dissertation at ESSLLI 2010 in Copenhagen. In Fall 2010, I spent two weeks at the Institute for Logic, Language, and Computation in Amsterdam at Johan's invitation, when I made substantial progress on the dissertation and presented a precursor of Chapter 2 at the University of Groningen. At ESSLLI 2011 in Ljubljana, I met with Johan to discuss a draft of part of the dissertation, and I was then well on my way. With this background, the process of writing the dissertation ran smoothly in Menlo Park.

None of this would have been possible without the extraordinary love and support of my wife, Allison, who has been an unfailing motivator through the long haul of graduate school, not to mention an inspiring educator and student herself. I do not know how she does it all, but I hope she knows how deeply I appreciate her.

Nor would it have been possible without the enormous strength and encouragement of my parents, Pamela and Dennis, and my brother, Taylor. I cannot express how much I owe to their years of positive influence. In the intellectual journey of earning a PhD, I have learned more about the adventure of a difficult research problem from my father than from anyone else, and I dedicate this dissertation to him.

Stanford, California
June 2012

Preface to the ILLC Version

I am grateful to my advisor, Johan van Benthem, for inviting me to submit my Stanford University dissertation to the ILLC Dissertation Series, where so many dissertations that I admire have appeared. I submitted the original version of this dissertation to Stanford in June 2012. For the ILLC version, I have made some corrections, included some proofs that were originally omitted, and added some footnotes with references to more recent work. I have resisted the urge to make more substantive changes based on subsequent developments. As of December 2013, the most recent formulations of the main ideas in the dissertation appear in Holliday 2013a,b,c.

Berkeley, California
December 2013

Contents

List of Tables

List of Figures

1

Introduction

> *If knowledge required the elimination of all logically possible alternatives, there would be no knowledge (at least of contingent truths).*
>
> – Alvin I. Goldman [1976, 775]

> *There are always, it seems, possibilities that our evidence is powerless to eliminate... If knowledge...requires the elimination of all competing possibilities (possibilities that contrast with what is known), then, clearly we seldom, if ever, satisfy the conditions for applying the concept.*
>
> – Fred I. Dretske [1981, 365]

Epistemic closure has been the subject of "one of the most significant disputes in epistemology over the last forty years" [Kvanvig, 2006, 256]. The starting point of the dispute is typically some version of the claim that knowledge is *closed under known implication* (see Dretske 2005). At its simplest, this is the claim that if an agent knows φ and knows that φ implies ψ, then the agent knows ψ:

$$(K\varphi \wedge K(\varphi \rightarrow \psi)) \rightarrow K\psi$$

in the language of basic epistemic logic.

An obvious objection to the simple version of closure under known implication is that an agent with bounded rationality may know φ and know that φ implies ψ, without "putting two and two together" and drawing a conclusion about ψ. Such an agent may not even believe ψ, let alone know it. The challenge of the much-discussed "problem of logical omniscience" (see, e.g., Stalnaker 1991, Halpern and Pucella 2011) is to develop a good theoretical model of the knowledge of such agents.

According to a different objection, made famous in epistemology by Dretske [1970] and Nozick [1981] (and applicable to more sophisticated closure claims), knowledge would not be closed under known implication even for "ideally astute logicians" [Dretske, 1970, 1010] who always put two and two together and believe all the consequences of what they know, based on what they know. This objection, rather than the logical omniscience objection, will be a focal point of what follows.[1]

One way to see the problem of epistemic closure is as a tension between closure under known implication and a widely held kind of *fallibilism* about knowledge. As the quotations from Goldman and Dretske at the beginning suggest, many epistemologists agree that if it were a general requirement on knowing a contingent truth φ that one rule out *every last way* in which φ could be false, then there would be little or no knowledge of contingent truths. As Lewis [1996, 549] colorfully explains:

> Let your paranoid fantasies rip—CIA plots, hallucinogens in the tap water, conspiracies to deceive, old Nick himself—and soon you find that uneliminated possibilities of error are everywhere. Those possibilities of error are far-fetched, of course, but possibilities all the same. They bite into even our most everyday knowledge. We never have infallible knowledge.

Fallibilism, in the Lewisian sense of the term, is the view that agents can know truths about the world even if they never rule out *all* possibilities of error or deception. To build up to the tension between fallibilism and closure, let us consider two examples.

Example 1.1 (Medical Diagnosis)**.** Two medical students, `A` and `B`, are subjected to a test. Their professor introduces them to the same patient, who presents various

[1] Other epistemologists who have denied closure under known implication in the relevant sense include McGinn [1984], Goldman [1986], Audi [1988], Heller [1999a], Harman and Sherman [2004, 2011], Lawlor [2005], Becker [2007], and Adams et al. [2012].

symptoms, and they are to make a diagnosis of the patient's condition. After some independent investigation, A and B conclude that the patient has a common condition c. In fact, they are correct. Yet only student A passes the test. For the professor wished to see if the students would check for another common condition c' that causes the same visible symptoms as c. While A ran laboratory tests to rule out c' before making the diagnosis of c, B made the diagnosis of c after only a physical exam.

In evaluating the students, the professor concludes that although both gave the correct diagnosis of c, student B did not know that the patient's condition was c, since B did not rule out the alternative of c'. Had the patient's condition been c', B might still have made the diagnosis of c, since the physical exam would not have revealed a difference. Student B was *lucky*. The condition B associated with the patient's visible symptoms happened to be the condition the patient had, but if the professor had chosen a patient with c', student B might have made a misdiagnosis. By contrast, student A secured against this possibility of error by running the lab tests. For this reason, the professor judges that A knew the patient's condition and passed the test.

Of course, A did not secure against *every* possibility of error. Suppose there is an extremely rare disease[2] x such that people with x appear to have c on lab tests given for c and c', even though people with x are *immune* to c, and only extensive further testing can detect x in its early stages. Should we say that A did not know that the patient had c after all, since A did not rule out x? As a fallibilist, the professor recognizes that the requirement that one rule out *all* possibilities of error would make knowledge impossible, since there are always some possibilities of error—however remote and far-fetched—that are uneliminated by one's evidence and experience. Yet if no one had any special reason to think that the patient may have had x instead of c, then it should not have been necessary to rule out such a remote possibility in order to know that the patient has the common condition (cf. Austin 1946, 156ff).

The second example is from Dretske 1981, 368-370, which I will quote at length.

Example 1.2 (Bird Watching)**.** "An amateur bird watcher spots a duck on his favorite Wisconsin pond. He quickly notes its familiar silhouette and markings and makes a

[2]Perhaps it has never been documented, but it is a possibility of medical theory.

mental note to tell his friends that he saw a Gadwall, a rather unusual bird in that part of the midwest. Since the Gadwall has a distinctive set of markings (black rump, white patch on the hind edge of the wing, etc.), markings that no other North American duck exhibits, and these markings were all perfectly visible, it seems reasonable enough to say that the bird-watcher *knows* that yonder bird is a Gadwall. He can see that it is.

Nevertheless, a concerned ornithologist is poking around in the vicinity, not far from where our bird-watcher spotted his Gadwall, looking for some trace of Siberian Grebes. Grebes are duck-like water birds, and the Siberian version of this creature is, when it is in the water, very hard to distinguish from a Gadwall duck. Accurate identification requires seeing the birds in flight since the Gadwall has a white belly and the Grebe a red belly—features that are not visible when the birds are in the water. The ornithologist has a hypothesis that some Siberian Grebes have been migrating to the midwest from their home in Siberia, and he and his research assistants are combing the midwest in search of confirmation.

Once we embellish our simple story in this way, intuitions start to diverge on whether our amateur bird-watcher does indeed know that yonder bird is a Gadwall duck (we are assuming, of course, that it *is* a Gadwall). Most people (I assume) would say that he did *not* know the bird to be a Gadwall if there actually were Siberian Grebes in the vicinity.... But what if the ornithologist's suspicions are unfounded. None of the Grebes have migrated. Does the bird-watcher still not know what he takes himself to know. Is, then, the simple presence of an ornithologist, with his false hypothesis, enough to rob the bird-watcher of his knowledge that the bird on the pond is a Gadwall duck? What if we suppose that the Siberian Grebes, because of certain geographical barriers, *cannot* migrate. Or suppose that there really are no Siberian Grebes—the existence of such a bird being a delusion of a crackpot ornithologist. We may even suppose that, in addition to there being no grebes, there is no ornithologist of the sort I described, but that people in the area believe that there is.... Or, finally, though no one believes any of this, some of the locals are interested in whether or not our birdwatcher knows that there are no look-alike migrant grebes in the area.

Somewhere in this progression philosophers, most of them anyway, will dig in their heels and say that the bird-watcher really does know that the bird he sees is a

Gadwall, and that he knows this despite his inability to justifiably rule out certain alternative possibilities.... He needn't be able to rule out the possibility that there are, somewhere in the world, look-alike grebes that have migrated to the midwest in order to know that the bird he saw was a Gadwall duck. These other possibilities are (whether the bird-watcher realizes it or not) simply too remote.

Most philosophers will dig in their heels here because they realize that if they don't, they are on the slippery slope to skepticism with nothing left to hang onto."

We can now pinpoint the tension between the fallibilist conclusions at the end of the examples and closure under known implication. In the case of Example 1.1, suppose for ease of exposition that student `A` is an ideally astute logician as described above. Further suppose that she knows that if her patient has c, then he does not have x (because x confers immunity to c), which we write as

(1) $K\,(c \rightarrow \neg x)$.[3]

Since `A` did not run any of the tests that could detect the presence or absence of x, arguably she does not know that the patient does not have x,

(2) $\neg K \neg x$.

Given the professor's judgment that `A` knows that the patient has condition c,

(3) Kc,

together (1) - (3) violate the following instance of closure under known implication:

(4) $(Kc \wedge K\,(c \rightarrow \neg x)) \rightarrow K \neg x$.

To maintain (4), one must say either that `A` does not know that the patient has condition c after all (having not excluded x), or else that `A` can know that a patient does not have a disease x without running any of the specialized tests for the disease (having learned instead that the patient has c, but from lab results consistent with x).[4] While

[3]For convenience, I use 'c', 'c'', and 'x' not only as names of conditions, but also as symbols for atomic sentences with the obvious intended meanings—that the patient has condition c, c', and x, respectively. Also for convenience, I will not add quotes when mentioning symbolic expressions.

[4]I have presented the argument so far without mentioning the concerns of epistemic *contextualists*. However, I will discuss contextualism at length in later chatpers, starting in §2.11.

the second option leads to what I will call the Problem of Vacuous Knowledge (or else the Problem of Knowledge Inflation), the first option leads to Radical Skepticism about knowledge, given the inevitability of uneliminated possibilities of error. We are lead to the same dilemma when we try to apply closure in Example 1.2. Either way, closure under known implication leads to problems for fallibilism.

This dissertation is a study of epistemic closure for fallibilists. As will quickly become apparent, in this study of closure I use epistemic-logical (especially model-theoretic) methods extensively, in addition to traditional epistemological ones. For modal logicians, I hope this study presents epistemology as an area of sophisticated theorizing in which modal-logical tools can help to clarify and systematize parts of the philosophical landscape. Doing so also benefits modal logic by broadening its scope, bringing interesting new structures and systems under its purview.

I conclude this brief introduction with an overview of the chapters to follow:

- Chapter 2 formalizes a family of relevant alternatives (RA) and subjunctivist theories of knowledge. The main result, the Closure Theorem, completely characterizes the valid epistemic closure principles according to these theories.

- Chapter 3 presents a unifying framework in which all of the RA and subjunctivist theories from Chapter 2 fit as special cases of what I call Fallibilism 1.0. The new framework also allows a finer-grained analysis of closure properties.

- Chapter 4 argues that any theory developed in the framework of Fallibilism 1.0 faces one of three serious problems, dubbed the Problem of Vacuous Knowledge, the Problem of Containment, and the Problem of Knowledge Inflation.

- Chapter 5 proposes a new framework for Fallibilism 2.0 based on what I call the Multipath Picture of Knowledge and the Transfer Picture of Deduction. These new pictures solve the problems raised in Chapter 4 for Fallibilism 1.0.

- Chapter 6 answers several objections according to which my fallibilist position on epistemic closure from Chapter 5—endorsing single-premise but not multi-premise logical closure—admits either *not enough* closure or *too much*.

2

Relevant Alternatives and Subjunctivism

In Chapter 1, I took the starting point of the debate over epistemic closure to be the principle of closure under known implication, written in its simplest form as $(K\varphi \wedge K(\varphi \to \psi)) \to K\psi$ in the language of basic epistemic logic. As noted, Dretske [1970], Nozick [1981] and others have rejected the validity of this principle on fallibilist grounds, even when applied only to ideally astute logicians who always "put two and two together" and believe all the consequences of what they know, based on what they know. In this chapter, we will analyze why closure under known implication fails according to their theories of knowledge, using Example 1.1 as our case study.

The closure of knowledge under known implication, henceforth referred to as 'K' after the modal axiom given above, is one closure principle among infinitely many. Although Dretske [1970] denied K, he accepted other closure principles, such as closure under conjunction elimination, $K(\varphi \wedge \psi) \to K\varphi$, and closure under disjunction introduction, $K\varphi \to K(\varphi \vee \psi)$ (1009). By contrast, Nozick [1981] was prepared to give up the first of those principles (228), although not the second (230n64, 692).

Dretske and Nozick not only provided examples in which they claimed K fails, but also proposed theories of knowledge that they claimed would explain the failures, as discussed below. Given such a theory, one may ask: is the theory committed to the failure of other, weaker closure principles, such as those mentioned above? Is it

committed to closure failures in situations other than those originally envisioned as counterexamples to K? The concern is that closure failures may spread, and they may spread to where no one wants them.

Pressing such a Problem of Containment has an advantage over other approaches to the debate over K. It appeals to considerations that both sides of the debate are likely to accept, rather than merely insisting on the plausibility of K (or of one of its more sophisticated versions). A clear illustration of this approach is Kripke's [2011] barrage of examples and arguments to the effect that closure failures are ubiquitous given Nozick's theory of knowledge. In a different way, Hawthorne [2004a, 41] presses the first part of the containment problem against Dretske and Nozick, as I critically discuss in §6.1.2.[1]

In this chapter, I formally assess the problem of containment for a family of prominent "modal" theories of knowledge (see, e.g., Pritchard 2008, Black 2010). In particular, I introduce formal models of the following: the *relevant alternatives* (RA) theories of Lewis [1996] and Heller [1989, 1999a]; one way of developing the RA theory of Dretske [1981] (based on Heller); the basic *tracking* theory of Nozick [1981]; and the basic *safety* theory of Sosa [1999]. A common feature of the theories of Heller, Nozick, and Sosa, which they share with those of Dretske [1971], Goldman [1976], and others, is some subjunctive or counterfactual-like condition(s) on knowledge, relating what an agent knows to what holds in selected *counterfactual possibilities* or *epistemic alternatives*.

Vogel [2007] characterizes *subjunctivism* as "the doctrine that what is distinctive about knowledge is essentially modal in character, and thus is captured by certain subjunctive conditionals" (73), and some versions of the RA theory have a similar flavor.[2] I will call this family of theories *subjunctivist flavored*. Reflecting their commonality,

[1] Lawlor [2005, 44] makes the methodological point about the advantage of raising the containment problem. It is noteworthy that Hawthorne takes a kind of proof-theoretic approach; he argues that a certain set of closure principles, not including K, suffices to derive the consequences that those who deny K wish to avoid, in which case they must give up a principle in the set. By contrast, our approach will be model-theoretic; we will study models of particular theories to identify those structural features that lead to closure failures.

[2] The view that knowledge has a modal character and the view that it is captured by subjunctive conditionals are different views. For example, Lewis [1996] adopts the modal view but not the subjunctive view. For more on subjunctivism, see Comesaña 2007.

my formal framework is based on the formal semantics for subjunctive conditionals in the style of Lewis 1973 and Stalnaker 1968. As a result, the epistemic logics studied here behave very differently than traditional epistemic logics in the style of Hintikka 1962 (Appendix §2.A contains a comparison).

One of the main result of this chapter is an exact characterization in propositional epistemic logic of the closure properties of knowledge according to the RA, tracking, and safety theories, as formalized. Below I preview some of the epistemological and logical highlights of this and other results. In later chapters, I further discuss the epistemological repercussions of these results.

Epistemological points. The extent to which subjunctivist-flavored theories of knowledge preserve closure is currently a topic of active discussion (see, e.g., Alspector-Kelly 2011, Adams et al. 2012). I show (in §2.6) that in contrast to Lewis's (non-subjunctive) theory, the other RA, tracking, and safety theories cited suffer from essentially the same widespread closure failures, far beyond the failure of K, which few if any proponents of these theories would welcome.[3] The theories' structural features responsible for these closure failures also lead (in §2.9) to serious problems of higher-order knowledge, including the possibility of knowing Fitch-paradoxical propositions [Fitch, 1963].

Analysis of these results reveals (in §2.10) that two parameters of a modal theory of knowledge affect whether it preserves closure. Each parameter has two values, for four possible parameter settings with respect to which each theory can be classified (Table 2.2). Of the theories mentioned, only Lewis's, with its unique parameter setting, fully preserves closure for a fixed context. (In §2.9 I clarify an issue, raised by Williamson [2001, 2009], about whether Lewis's theory validates strong principles of higher-order knowledge.) Finally, I formalize Lewis's view of the dynamics of context change (in §2.11), leading to the result that for every closure principle that fails for the other theories with respect to a fixed context, an "inter-context" version of that

[3]While closure failures for these subjunctivist-flavored theories go too far in some directions, in other directions they do not go far enough for the purposes of Dretske and Nozick: all of these theories validate closure principles (see §2.6) that appear about as dangerous as K in arguments for radical skepticism about knowledge. This fact undermines the force of responding to skepticism by rejecting K on subjunctivist grounds, as Nozick does.

closure principle fails for Lewis's contextualist theory.

In the terminology of Dretske [1970], the knowledge operator for Lewis's theory is *fully penetrating* for a fixed context. For all of the other theories, the knowledge operator lacks the basic closure properties that Dretske wanted from a *semi-penetrating* operator. Contrary to common assumptions in the literature (perhaps due to neglect of the second theory parameter in §2.10), serious closure failures are not avoided by modified subjunctivist theories, such as DeRose's [1995] modified tracking theory or the modified safety theory with *bases*, which I treat formally in §2.D. For those seeking a balance of closure properties between full closure and not enough closure, it appears necessary to abandon essential elements of the standard theories. I will show how to do just that Chapter 5, but there is much ground to cover before then.

While I take the results of this chapter to be negative for subjunctivist-flavored theories qua theories of knowledge, we can also take them to be neutral results about other desirable epistemic properties, viz., the properties of having ruled out the relevant alternatives to a proposition, of having a belief that tracks the truth of a proposition, of having a safe belief in a proposition, etc., even if these are neither necessary nor sufficient for knowledge (see §2.6 and §2.8).

Logical points. This chapter demonstrates the effectiveness of an alternative approach to proving modal completeness theorems, illustrated by van Benthem [2010, §4.3] for the normal modal logic **K**, in a case that presents difficulties for a standard canonical model construction. The key element of the alternative approach is a "modal decomposition" result. By proving such results (Theorem 2.1), we will obtain completeness (Corollary 2.4) of two non-normal modal logics with respect to new semantics mixing elements of ordering semantics [Lewis, 1981] and relational semantics [Kripke, 1963]. One of these logics, dubbed *the logic of ranked relevant alternatives*, appears not to have been previously identified in the modal logic literature. Further results on decidability (Corollary 2.1), finite models (Corollary 2.2), and complexity (Corollary 2.3) follow from the proof of the modal decomposition results.

In addition to these technical points, this chapter—indeed, this dissertation—aims to show that for modal logicians, epistemology represents an area of sophisticated theorizing in which modal-logical tools can help to clarify and systematize parts of

the philosophical landscape. Doing so also benefits modal logic by broadening its scope, bringing interesting new structures and systems under its purview.

In §2.1, I begin by reviewing some relevant background of the epistemic closure debate. After introducing our formal epistemic language in §2.2, I introduce the formal framework for the study of closure in RA and subjunctivist theories in §2.4 and §2.5. With this setup, I state and prove the main theorems in §2.6 and §2.8, with an interlude on relations between RA and subjunctivist models in §2.7. Finally, I investigate higher-order knowledge in §2.9, discuss the relation between theory parameters and closure failures in §2.10, and model the dynamics of context in §2.11.

2.1 Background

In this chapter, we will use Example 1.1 as our running case study. Recall how the problem of closure arises in this case. As before, suppose that student `A` knows that if her patient has c, then he does not have x (because x confers immunity to c),

(1) $K\,(c \to \neg x)$.

Since `A` did not run any of the tests that could detect the presence or absence of x, arguably she does not know that the patient does not have x,

(2) $\neg K \neg x$.

Given the professor's judgment that `A` knows that the patient has condition c,

(3) Kc,

together (1) - (3) violate the following instance of closure under known implication:

(4) $(Kc \wedge K\,(c \to \neg x)) \to K \neg x$.

To maintain (4), one must say either that `A` does not know that the patient has condition c after all (having not excluded x), or else that `A` can know that a patient does not have a disease x without running any of the specialized tests for the disease (having learned instead that the patient has c, but from lab results consistent with x). While

the second option leads to what I will call the Problem of Vacuous Knowledge (or else the Problem of Knowledge Inflation), the first option leads to Radical Skepticism about knowledge, given the inevitability of uneliminated possibilities of error. Either way, closure under known implication leads to problems.

Dretske [1970] and Nozick [1981] propose to resolve the inconsistency of (1) - (4), a version of the now standard "skeptical paradox" [Cohen, 1988, DeRose, 1995], by denying the validity of K in general and its instance (4) in particular. This denial has nothing to do with the "putting two and two together" problem noted in §1. The claim is that K would fail even for Dretske's [1970] "ideally astute logicians" (1010). I will characterize an ideally astute logician (IAL) in terms of two properties.

- *Validity omniscience*: the IAL knows all classically valid logical principles.[4]
- *Full doxastic closure*: the IAL believes all the classical logical consequences of the set of propositions she believes.[5]

Dretske's explanation for why K fails even for such agents is in terms of the RA theory. (We turn to Nozick's view in §2.5.) For this theory, to know p is (to truly believe p and) to have *ruled out the relevant alternatives to* p. In coming to know c and $c \rightarrow \neg x$, the agent rules out certain relevant alternatives. In order to know $\neg x$, the agent must rule out certain relevant alternatives. But the relevant alternatives in the two cases *are not the same*. According to our earlier reasoning, x is not an alternative that must be ruled out in order for Kc to hold. But x *is* an alternative that must be ruled out in order for $K\neg x$ to hold (cf. Remark 2.3 in §2.4). It is because the relevant alternatives may be different for what is in the antecedent of K and what is in the consequent that instances like (iv) can fail.

In an influential objection to Dretske, Stine [1976] claimed that to allow for the relevant alternatives to be different for the premises and conclusion of an argument

[4]Note the distinction with a stronger property of *consequence omniscience* (standardly "logical omniscience"), that one knows all the logical consequences of what one knows.

[5]We may add that such an agent has come to believe these logical consequences by "competent deduction," rather than (only) by some other means, but we will not explicitly represent methods or bases of beliefs until §2.D (see Remark 2.1). By "all the logical consequences" I mean all of those *involving concepts that the agent grasps*. Otherwise one might believe p and yet fail to believe $p \vee q$ because one does not grasp q (see Williamson 2000, 283). Assume that the agent grasps all of the atomic $p, q, r, \ldots$ of Definition 2.1.

about knowledge "would be to commit some logical sin akin to equivocation" (256). Yet as Heller [1999a] points out in Dretske's defence, a similar charge of equivocation could be made (incorrectly) against accepted counterexamples to the principles of transitivity or antecedent strengthening for counterfactuals. If we take a counterfactual $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ to be true iff the "closest" φ-worlds are ψ-worlds, then the inference from $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ to $(\varphi \wedge \chi) \mathrel{\Box\!\!\rightarrow} \psi$ is invalid because the closest $(\varphi \wedge \chi)$-worlds may not be among the closest φ-worlds. Heller argues that there is no equivocation in such counterexamples since we use the same, fixed similarity ordering of worlds to evaluate the different conditionals. Similarly, in the example of closure failure, the most relevant $\neg c$-worlds may differ from the most relevant x-worlds—so one can rule out the former without ruling out the latter—even assuming a fixed relevance ordering of worlds. In this defense of Dretske, Heller brings the RA theory closer to subjunctivist theories that place counterfactual conditions on knowledge, an important theme that will return.

In another influential objection to Dretske (with origins in Stine 1976), Lewis [1996] and others [Cohen, 1988, DeRose, 1995] attempt to explain away apparent closure failures by appeal to *epistemic contextualism*, the thesis that the truth values of knowledge attributions are context sensitive. According to Lewis's contextualist RA theory, in the context $\mathcal{C}$ of our conversation before we raised the possibility of the rare disease x, that possibility was irrelevant; so although A had not eliminated the possibility of x, we could truly say in $\mathcal{C}$ that A knew (at time t) that the patient's condition was c (Kc). However, by raising the possibility of x in our conversation, we changed the context to a new $\mathcal{C}'$ in which the uneliminated possibility of x was relevant. Hence we could truly say in $\mathcal{C}'$ that A did *not* know that the patient did not have x ($\neg K \neg x$), although A knew that x confers immunity to c ($K(c \rightarrow \neg x)$), which did not require ruling out x. Is this a violation of K in context $\mathcal{C}'$? It is not, because in $\mathcal{C}'$, unlike $\mathcal{C}$, we could *no longer* truly say that A knew (at t) that the patient's condition was c (Kc), given that A had not eliminated the newly relevant possibility of x. Moreover, Lewis argues that there is no violation of K in context $\mathcal{C}$ either:

> Knowledge *is* closed under implication.... Implication preserves truth—that is, it preserves truth in any given, fixed context. But if we switch

> contexts, all bets are off.... Dretske gets the phenomenon right...it is just that he misclassifies what he sees. He thinks it is a phenomenon of logic, when really it is a phenomenon of pragmatics. Closure, rightly understood, survives the rest. If we evaluate the conclusion for truth not with respect to the context in which it was uttered, but instead with respect to the different context in which the premise was uttered, then truth is preserved. (564)

In other words, Lewis claims that if we evaluate the consequent of (4), $K\neg x$, with respect to the context $\mathcal{C}$ of our conversation before we raised the possibility of x, then it should come out *true*—despite the fact that `A` had not eliminated the possibility of x through any special tests—because the possibility of x was irrelevant in $\mathcal{C}$. If this is correct, then there is no violation of K in either context $\mathcal{C}'$ or $\mathcal{C}$. But how significant is it to "preserve closure" for a fixed context by claiming that when we try to reason in real time with K, we end up changing the context in the process so K does not apply? This is a question that bothered Dretske [2005], and we will return to it later.

In addition to the two objections to Dretske stated above, there is the worry stated at the beginning—that denying closure under known implication may commit one to giving up other, weaker closure principles, what I called the Problem of Containment. In what follows, I develop a formal framework to study these issues systematically.

2.2 Epistemic Language

With the background of §2.1, let us formulate the question of closure to be studied. We begin with the official definition of our (first) propositional epistemic language. The framework of §2.4 - 2.5 could be extended for quantified epistemic logic, but there is already plenty to investigate in the propositional case.[6]

[6]It is not difficult to extend the framework of §2.4 - 2.5 to study closure principles of the form shown below Definition 2.1 where the φ's and ψ's may contain first-order quantifiers, provided that no free variables are allowed within the scope of any K operator. The closure behavior of K with respect to $\forall$ and $\exists$ can be anticipated on the basis of the closure behavior of K with respect to $\wedge$ and $\vee$ shown in Theorem 2.1. Of course, interesting complications arise whenever we allow quantification into the scope of a K operator (see Holliday and Perry 2013).

Definition 2.1 (Epistemic Language)**.** Let $\mathsf{At} = \{p, q, r, \dots\}$ be a countable set of atomic sentences. The *epistemic language* is defined inductively by

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K\varphi,$$

where $p \in \mathsf{At}$. As usual, expressions containing $\vee$, $\rightarrow$, and $\leftrightarrow$ are abbreviations, and by convention $\wedge$ and $\vee$ bind more strongly than $\rightarrow$ or $\leftrightarrow$ in the absence of parentheses; we take $\top$ to be an arbitrary tautology (e.g., $p \vee \neg p$), and $\bot$ to be $\neg\top$. The *modal degree* of a formula φ is defined recursively as follows:

$$\begin{aligned} d(p) &= 0 \\ d(\neg\varphi) &= d(\varphi) \\ d(\varphi \wedge \psi) &= \max(d(\varphi), d(\psi)) \\ d(K\varphi) &= d(\varphi) + 1. \end{aligned}$$

A formula φ is *propositional* iff $d(\varphi) = 0$ and *flat* iff $d(\varphi) \leq 1$.

The flat fragment has a special place in the study of closure, which need not involve higher-order knowledge. In the most basic case we are interested in whether for a valid propositional formula $\varphi_1 \wedge \dots \wedge \varphi_n \rightarrow \psi$, the associated "closure principle" $K\varphi_1 \wedge \dots \wedge K\varphi_n \rightarrow K\psi$ is valid, according to some semantics for the K operator. More generally, we will consider principles of the form $K\varphi_1 \wedge \dots \wedge K\varphi_n \rightarrow K\psi_1 \vee \dots \vee K\psi_m$, allowing each φ_i and ψ_j to be of arbitrary modal degree. As above, we ask whether such principles hold for ideally astute logicians. The question can be understood in two ways, depending on whether we have in mind what may be called *pure*, *empirical*, or *deductive* closure principles.

Remark 2.1 (Types of Closure)**.** For example, if we understand the principle $K(\varphi \wedge \psi) \rightarrow K\psi$ as a *pure* closure principle, then its validity means that an agent cannot know $\varphi \wedge \psi$ without knowing ψ—regardless of whether the agent came to believe ψ by "competent deduction" from $\varphi \wedge \psi$.[7] (Perhaps she came to believe ψ from perception,

[7] Harman and Sherman [2004] criticize Williamson's [2000] talk of "deduction" as extending knowledge for its "presupposition that deduction is a kind of inference, something one does" (495). I will return to this issue in §5.4.

φ from testimony, and $\varphi \wedge \psi$ by competent deduction from φ and ψ.) More generally, if we understand $K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi$ as a pure closure principle, its validity means that an agent cannot know $\varphi_1, \ldots, \varphi_n$ without knowing ψ. Understood as an *empirical* closure principle, its validity means that an agent who has done enough empirical investigation to know $\varphi_1, \ldots, \varphi_n$ has done enough to know ψ. Finally, understood as a *deductive* closure principle, its validity means that *if* the agent came to believe ψ from $\varphi_1, \ldots, \varphi_n$ by competent deduction, all the while knowing $\varphi_1, \ldots, \varphi_n$, then she knows ψ. As suggested by Williamson [2000, 282f], it is highly plausible that $K(\varphi \wedge \psi) \rightarrow K\psi$ is a pure (and hence empirical and deductive) closure principle. By contrast, closure under known implication is typically understood as only an empirical or deductive closure principle.[8] We will not explicitly explicitly represent in our language or models the idea of deductive closure until Appendix 2.D, where I formalize versions of the tracking and safety theories that take into account *methods* or *bases* of beliefs. It is first necessary to understand the structural reasons for why the basic RA, tracking, and safety conditions are not purely or empirically closed, in order to understand whether the refined theories solve all the problems of epistemic closure. As shown in §2.D, there are failures of pure and deductive closure for the tracking theory *with methods*, for the structural reasons identified here. The safety theory *with bases* arguably supports deductive closure, but also has problems with pure closure for the structural reasons identified here.

2.3 Three Distinctions

Before giving RA semantics for the epistemic language of Definition 2.1, let us observe several distinctions between different versions of the RA theory.

The first concerns the nature of the "alternatives" that one must rule out to know p. Are they *possibilities* (or ways the world could/might be) in which p is false?[9]

[8] Deductive closure principles belong to a more general category of "active" closure principles, which are conditional on the agent performing some action, of which deduction is one example. As Johan van Benthem (personal communication) suggests, the active analogue of K has the form $K\varphi \wedge K(\varphi \rightarrow \psi) \rightarrow [a]K\psi$, where $[a]$ stands for *after action a*.

[9] In order to deal with self-locating knowledge, one may take the alternatives to be "centered"

Or are they *propositions* that entail the negation of p? Both views are common in the literature, sometimes within a single author. Although earlier I wrote in a way suggestive of the second view, in what follows I adopt the first view, familiar in the epistemic logic tradition since Hintikka, since it fits the theories I will formalize. In §4.A, I show how to treat views that take alternatives to be propositions.

The second distinction concerns the structure of relevant alternatives. On one hand, Dretske [1981] states the following definition in developing his RA theory: "call the set of possible alternatives that a person must be in an evidential position to exclude (when he knows P) the *Relevancy Set* (RS)" (371). On the other hand, Heller [1999a] considers (and rejects) an interpretation of the RA theory in which "there is a certain set of worlds selected as relevant, and S must be able to rule out the not-p worlds within that set" (197).

According to Dretske, for every proposition P, there is a relevancy set for that P. Let us translate this into Heller's talk of worlds. Where $\overline{P}$ is the set of all worlds in which P is false, let $\mathsf{r}(P)$ be the relevancy set for P, so $\mathsf{r}(P) \subseteq \overline{P}$. To be more precise, since objective features of an agent's situation in world w may affect what alternatives are relevant and therefore what it takes to know P in w (see Dretske 1981, 377 and DeRose 2009, 30f on "subject factors"), let us write '$\mathsf{r}(P, w)$' for the relevancy set for P in world w, so $\mathsf{r}(P, w)$ may differ from $\mathsf{r}(P, v)$ for a distinct world v in which the agent's situation is different. Finally, if we allow (unlike Dretske) that the conversational context $\mathcal{C}$ of those attributing knowledge to the agent can also affect what alternatives are relevant in a given situation w and therefore what it takes to count as knowing P in w relative to $\mathcal{C}$ (see DeRose 2009, 30f on "attributor factors"), then we should write '$\mathsf{r}_{\mathcal{C}}(P, w)$' to make the relativization to context explicit.

The quote from Dretske suggests the following definition:

> According to a $\mathsf{RS}_{\forall\exists}$ theory, for every context $\mathcal{C}$, for every world w, and for every ($\forall$) proposition P, there is ($\exists$) a set of *relevant* (*in* w) *not-*P *worlds*, $\mathsf{r}_{\mathcal{C}}(P, w) \subseteq \overline{P}$, such that in order to know P in w (relative to $\mathcal{C}$) one must rule

worlds or possible individuals (see Lewis 1986, §1.4 and references therein). Another question is whether we should think of what is ruled out by knowledge as including *ways the world could not be* (metaphysically "impossible worlds" or even logically impossible worlds), in addition to *ways the world could be*. I discuss this question in §5.1.3.

> out the worlds in $\mathsf{r}_c(P, w)$.

By contrast, the quote from Heller suggests the following definition:

> According to a $\mathsf{RS}_{\exists\forall}$ theory, for every context $\mathcal{C}$ and for every world w, there is ($\exists$) a set of *relevant* (*in* w) *worlds*, $\mathsf{R}_c(w)$, such that for every ($\forall$) proposition P, in order to know P in w (relative to $\mathcal{C}$) one must rule out the not-P worlds in that set, i.e., the worlds in $\mathsf{R}_c(w) \cap \overline{P}$.

As a simple logical observation, every $\mathsf{RS}_{\exists\forall}$ theory is a $\mathsf{RS}_{\forall\exists}$ theory (take $\mathsf{r}_c(P, w) = \mathsf{R}_c(w) \cap \overline{P}$), but not necessarily *vice versa*. From now on, when I refer to $\mathsf{RS}_{\forall\exists}$ theories, I have in mind theories that are not also $\mathsf{RS}_{\exists\forall}$ theories. This distinction is at the heart of the disagreement about epistemic closure between Dretske and Lewis [1996], as Lewis clearly adopts an $\mathsf{RS}_{\exists\forall}$ theory.

In a *contextualist* $\mathsf{RS}_{\exists\forall}$ theory, such as Lewis's, the set of relevant worlds may change as context changes. Still, for any given context $\mathcal{C}$, there is a set $\mathsf{R}_c(w)$ of relevant (at w) worlds, which does not depend on the particular proposition in question. The $\mathsf{RS}_{\forall\exists}$ vs. $\mathsf{RS}_{\exists\forall}$ distinction is about how theories view the relevant alternatives *with respect to a fixed context*. In the following sections we will study which closure principles hold for different theories with respect to a fixed context. In §2.11, we will extend the framework to context change.

A third distinction between versions of the RA theory concerns different notions of ruling out or eliminating alternatives (possibilities or propositions). On one hand, Lewis [1996] proposes that "a possibility . . . [v] . . . is *uneliminated* iff the subject's perceptual experience and memory in . . . [v] . . . exactly match his perceptual experience and memory in actuality" (553). On the other hand, Heller [1999a] proposes that "S's ability to rule out not-p be understood thus: S does not believe p in any of the relevant not-p worlds" (98). First, we model the RA theory with a Lewis-style notion of elimination. By 'Lewis-style', I do not mean a notion that involves experience or memory; I mean any notion of elimination that allows us to decide whether a possibility v is eliminated by an agent in w *independently* of any proposition P in question, as Lewis's notion does. In §2.5, we turn to Heller's notion, which is closely related to Nozick's [1981] tracking theory. We compare the two notions in §2.10.

2.4 Relevant Alternatives (RA) Models

In this section we define our first class of models, following Heller's RA picture of "worlds surrounding the actual world ordered according to how realistic they are, so that those worlds that are more realistic are closer to the actual world than the less realistic ones" [1989, 25] with "those that are too far away from the actual world being irrelevant" [1999a, 199]. These models represent the epistemic state of an agent from a third-person perspective. We should not assume that anything in the model is something that the agent has in mind. Contextualists should think of the model $\mathcal{M}$ as associated with a fixed context of knowledge attribution, so a change in context corresponds to a change in models from $\mathcal{M}$ to $\mathcal{M}'$, an idea developed formally in §2.11. Just as the model is not something that the agent has in mind, it is not something that particular speakers attributing knowledge to the agent have in mind either. For possibilities may be relevant and hence should be included in our model, even if the attributors are not considering them (see DeRose 2009, 33).

Finally, for simplicity (and in line with Lewis 1996) we will not represent in our RA models an agent's beliefs separately from her knowledge. Adding the doxastic machinery of §2.5 (which guarantees doxastic closure) would be easy, but if the only point were to add *believing* φ as a necessary condition for knowing φ, this would not change any of our results about RA knowledge.[10]

Definition 2.2 (RA Model)**.** A *relevant alternatives model* is a tuple $\mathcal{M}$ of the form $\langle W, \rightarrow\!\!\!\!\triangleright, \preceq, V\rangle$ where:

1. W is a non-empty set;
2. $\rightarrow\!\!\!\!\triangleright$ is a reflexive binary relation on W;
3. $\preceq$ assigns to each $w \in W$ a binary relation $\preceq_w$ on some $W_w \subseteq W$;
   (a) $\preceq_w$ is reflexive and transitive;

[10] If one were to also adopt a variant of Lewis's [1996] *Rule of Belief* according to which any world v doxastically accessible for the agent in w must be relevant and uneliminated for the agent in w (i.e., using notation introduced below, wDv implies $v \in \mathrm{Min}_{\preceq_w}(W)$ and $w \rightarrow\!\!\!\!\triangleright\, v$), then belief would already follow from the knowledge condition of Definition 2.4.

(b) $w \in W_w$, and for all $v \in W_w$, $w \preceq_w v$;

4. V assigns to each $p \in \mathsf{At}$ a set $V(p) \subseteq W$.

For $w \in W$, the pair $\mathcal{M}, w$ is a *pointed* model.

I refer to elements of W as "worlds" or "possibilities" interchangeably.[11] As usual, think of $V(p)$ as the set of worlds where the atomic sentence p holds.

Take $w \rightarrowtriangle v$ to mean that v is an *uneliminated* possibility for the agent in w.[12] For generality, I assume only that $\rightarrowtriangle$ is reflexive, reflecting the fact that an agent can never eliminate her actual world as a possibility. According to Lewis's [1996] notion of elimination, $\rightarrowtriangle$ is an equivalence relation. However, whether we assume transitivity and symmetry in addition to reflexivity does not affect our main results (see Remark 2.8). This choice only matters if we make further assumptions about the $\preceq_w$ relations, which we discuss in §2.9.

Take $u \preceq_w v$ to mean that u is *at least as relevant* (at w) as v is.[13] A relation satisfying Definition 2.2.3a is a *preorder*. The family of preorders in an RA model is like one of Lewis's (weakly centered) comparative similarity systems [1973, §2.3] or standard γ-models [1971], but without his assumption that each $\preceq_w$ is *total* on its field W_w (see Definition 2.3.3). Condition 3b, that w is at least as relevant at w as any other world is, corresponds to Lewis's [1996] *Rule of Actuality*, that "actuality is always a relevant alternative" (554).

By allowing $\preceq_w$ and $\preceq_v$ to be different for distinct worlds w and v, we allow the world-relativity of comparative relevance (based on differences in "subject factors")

[11] Lewis [1996] is neutral on whether the *possibilities* referred to in his definition of knowledge must be "maximally specific" (552), as *worlds* are often thought to be. It should be clear that the examples in this chapter do not depend on taking possibilities to be maximally specific either.

[12] Those who have used standard Kripke models for epistemic modeling should note an important difference in how we use W and $\rightarrowtriangle$. We include in W not only possibilities that the agent has not eliminated, but also possibilities that the agent *has* eliminated, including possibilities v such that $w \not\rightarrowtriangle v$ for all w distinct from v. While in standard Kripke semantics for the (single-agent) epistemic language, such a possibility v can always be deleted from W without changing the truth value of any formula at w (given the invariance of truth under $\rightarrowtriangle$-generated submodels), this will *not* be the case for one of our semantics below (D-semantics). So if we want to indicate that an agent in w has eliminated a possibility v, we do not leave it out of W; instead, we include it in W and set $w \not\rightarrowtriangle v$.

[13] One might expect $u \preceq_w v$ to mean that v is at least as relevant (at w) as u is, by analogy with $x \leq y$ in arithmetic, but Lewis's [1973, §2.3] convention is now standard.

discussed above. A fixed context may help to determine not only which possibilities are relevant, given the way things actually are, but also which possibilities would be relevant were things different. Importantly, we also allow $\preceq_w$ and $\preceq_v$ to be different when v is an uneliminated possibility for the agent in w, so $w \rightarrowtriangle v$. In other words, we do not assume that in w the agent can eliminate any v for which $\preceq_v \neq \preceq_w$. As Lewis [1996] put it, "the subject himself may not be able to tell what is properly ignored" (554). We will return to these points in §2.9 in our discussion of higher-order knowledge.

Notation 2.1 (Derived Relations, Min)**.** Where $w, v, u \in W$ and $S \subseteq W$,

- $u \prec_w v$ iff $u \preceq_w v$ and not $v \preceq_w u$; and $u \simeq_w v$ iff $u \preceq_w v$ and $v \preceq_w u$;
- $\mathrm{Min}_{\preceq_w}(S) = \{v \in S \cap W_w \mid \text{there is no } u \in S \text{ such that } u \prec_w v\}$.

Hence $u \prec_w v$ means that possibility u is *more relevant* (at w) than possibility v is, while $u \simeq_w v$ means that they are equally relevant. $\mathrm{Min}_{\preceq_w}(S)$ is the set of *most relevant* (at w) possibilities out of those in S that are ordered by $\preceq_w$, in the sense that there are no other possibilities that are more relevant (at w).

Definition 2.3 (Types of Orderings)**.** Consider an RA model $\mathcal{M} = \langle W, \rightarrowtriangle, \preceq, V\rangle$ with $w \in W$.

1. $\preceq_w$ is *well-founded* iff for every non-empty $S \subseteq W_w$, $\mathrm{Min}_{\preceq_w}(S) \neq \emptyset$;
2. $\preceq_w$ is *linear* iff for all $u, v \in W_w$, either $u \prec_w v$, $v \prec_w u$, or $u = v$;
3. $\preceq_w$ is *total* iff for all $u, v \in W_w$, $u \preceq_w v$ or $v \preceq_w u$;
4. $\preceq_w$ has a *universal field* iff $W_w = W$;
5. $\preceq_w$ is *centered* (*weakly centered*) iff $\mathrm{Min}_{\preceq_w}(W) = \{w\}$ ($w \in \mathrm{Min}_{\preceq_w}(W)$).

If a property holds of $\preceq_v$ for all $v \in W$, then we say that $\mathcal{M}$ has the property.

Well-foundedness is a (language-independent) version of the "Limit Assumption" discussed by Lewis [1973, §1.4]. Together well-foundedness and linearity amount to

"Stalnaker's Assumption" (ibid., §3.4). Totality says that any worlds in the field of $\preceq_w$ are comparable in relevance. So a total preorder $\preceq_w$ is a relevance *ranking* of worlds in W_w. Universality (ibid., §5.1) says that all worlds are assessed for relevance at w. Finally, (with Def. 2.2.3b) centering (ibid., §1.3) says that w is *the* most relevant world at w, while weak centering (ibid., §1.7) (implied by Def. 2.2.3b) says that w is *among* the most relevant.

I assume well-foundedness (always satisfied in finite models) in what follows, since it allows us to state more perspicuous truth definitions. However, this assumption does not affect our results (see Remark 2.7). By contrast, totality does make a difference in valid closure principles for one of our theories (see Fact 2.6), while the addition of universality does not (see Prop. 2.3). I comment on linearity and centering vs. weak centering after Definition 2.5.

We now interpret the epistemic language of Definition 2.1 in RA models, considering three semantics for the K operator. I call these C-semantics, for **C**artesian, D-semantics, for **D**retske, and L-semantics, for **L**ewis. C-semantics is not intended to capture Descartes' view of knowledge. Rather, it is supposed to reflect a high standard for the truth of knowledge claims—knowledge requires ruling out all possibilities of error, however remote—in the spirit of Descartes' worries about error in the First Meditation; formally, C-semantics is just the standard semantics for epistemic logic in the tradition of Hintikka [1962], but I reserve 'H-semantics' for later. D-semantics is one way (but not the only way) of understanding Dretske's [1981] $\mathsf{RS}_{\forall\exists}$ theory, using Heller's [1989, 1999a] picture of relevance orderings over possibilities.[14] Finally, L-semantics follows Lewis's [1996] $\mathsf{RS}_{\exists\forall}$ theory (for a fixed context).

Definition 2.4 (Truth in an RA Model)**.** Given a well-founded RA model $\mathcal{M} = \langle W, \rightarrow\!\!\triangleright, \preceq, V\rangle$ with $w \in W$ and a formula φ in the epistemic language, we define

[14]Later I argue that there is a better way of understanding Dretske's [1981] $\mathsf{RS}_{\forall\exists}$ theory, without the familiar world-ordering picture. Hence I take the 'D' for D-semantics as loosely as the 'C' for C-semantics. Nonetheless, it is a helpful mnemonic for remembering that D-semantics formalizes an RA theory that allows closure failure, as Dretske's does, while L-semantics formalizes an RA theory that does not, like Lewis's.

$\mathcal{M}, w \vDash_x \varphi$ (φ is true at w in $\mathcal{M}$ according to X-semantics) as follows:

$$\begin{array}{lll} \mathcal{M}, w \vDash_x p & \text{iff} & w \in V(p); \\ \mathcal{M}, w \vDash_x \neg\varphi & \text{iff} & \mathcal{M}, w \nvDash_x \varphi; \\ \mathcal{M}, w \vDash_x (\varphi \wedge \psi) & \text{iff} & \mathcal{M}, w \vDash_x \varphi \text{ and } \mathcal{M}, w \vDash_x \psi. \end{array}$$

For the K operator, the C-semantics clause is that of standard modal logic:

$$\mathcal{M}, w \vDash_c K\varphi \text{ iff } \forall v \in W\text{: if } w \rightarrowtriangle v \text{ then } \mathcal{M}, v \vDash_c \varphi,$$

which states that φ is known at w iff φ is true in all possibilities uneliminated at w. I will write this clause in another, equivalent way below, for comparison with the D- and L-semantics clauses. First, we need two pieces of notation.

Notation 2.2 (Extension and Complement)**.** Where $\mathcal{M} = \langle W, \rightarrowtriangle, \preceq, V\rangle$,

- $[\![\varphi]\!]^{\mathcal{M}}_x = \{v \in W \mid \mathcal{M}, v \vDash_x \varphi\}$ is the set of worlds where φ is true in $\mathcal{M}$ according to X-semantics; if $\mathcal{M}$ and x are clear from context, I write '$[\![\varphi]\!]$'.
- For $S \subseteq W$, $\overline{S} = \{v \in W \mid v \notin S\}$ is the complement of S in W. When W may not be clear from context, I write '$W \setminus S$' instead of '$\overline{S}$'.

Definition 2.5 (Truth in an RA Model cont.)**.** For C-, D-, and L-semantics, the clauses for the K operator are:[15]

$$\begin{array}{lll} \mathcal{M}, w \vDash_c K\varphi & \text{iff} & \forall v \in \overline{[\![\varphi]\!]}_c\text{: } w \not\rightarrowtriangle v; \\ \mathcal{M}, w \vDash_d K\varphi & \text{iff} & \forall v \in \mathrm{Min}_{\preceq_w}\big(\overline{[\![\varphi]\!]}_d\big)\text{: } w \not\rightarrowtriangle v; \\ \mathcal{M}, w \vDash_l K\varphi & \text{iff} & \forall v \in \mathrm{Min}_{\preceq_w}(W) \cap \overline{[\![\varphi]\!]}_l\text{: } w \not\rightarrowtriangle v. \end{array}$$

According to C-semantics, in order for an agent to know φ in world w, *all* of the $\neg\varphi$-possibilities must be eliminated by the agent in w. According to D-semantics, for

[15] Instead of thinking in terms of three different satisfaction relations, $\vDash_c$, $\vDash_d$, and $\vDash_l$, some readers may prefer to think in terms of one satisfaction relation, $\vDash$, and three different operators, K_c, K_d, and K_l. I choose to subscript the turnstile instead of the operator in order to avoid proliferating subscripts in formulas. One should not read anything more into this practical choice of notation. (However, note that epistemologists typically take themselves to be proposing different accounts of the conditions under which an agent has knowledge, rather than proposing different epistemic notions of knowledge$_1$, knowledge$_2$, etc.)

any φ there is a set $\mathrm{Min}_{\preceq_w}\big(\overline{[\![\varphi]\!]}_d\big)$ of *most relevant (at w)* $\neg\varphi$-possibilities that the agent must eliminate in order to know φ. Finally, according to L-semantics, there is a set of relevant possibilities, $\mathrm{Min}_{\preceq_w}(W)$, such that for any φ, in order to know φ the agent must eliminate the $\neg\varphi$-possibilities *within that set*. Recall the $\mathsf{RS}_{\forall\exists}$ vs. $\mathsf{RS}_{\exists\forall}$ distinction above.

If φ is true at all pointed models according to X-semantics, then φ is *X-valid*, written '$\vDash_x \varphi$'. Since the semantics do not differ with respect to propositional formulas φ, I sometimes omit the subscript in '$\vDash_x$' and simply write '$\mathcal{M}, w \vDash \varphi$'. These conventions also apply to the semantics in Definition 2.7.

Since for L-semantics we think of $\mathrm{Min}_{\preceq_w}(W)$ as the set of simply *relevant* worlds, ignoring the rest of $\preceq_w$, we allow $\mathrm{Min}_{\preceq_w}(W)$ to contain multiple worlds. Hence with L-semantics we assume neither centering nor linearity, which implies centering by Definition 2.2.3b. For D-semantics, whether we assume centering/linearity does not affect our results (as shown in §2.6.2).

It is easy to check that according to C/D/L-semantics, whatever is known is true. For D- and L-semantics, Fact 2.1 reflects Lewis's [1996, 554] observation that the veridicality of knowledge follows from his Rule of Actuality, given that an agent can never eliminate her actual world as a possibility. Formally, veridicality follows from the fact that w is minimal in $\preceq_w$ and $w \twoheadrightarrow w$.

Fact 2.1 (Veridicality)**.** $K\varphi \to \varphi$ is C/D/L-valid.

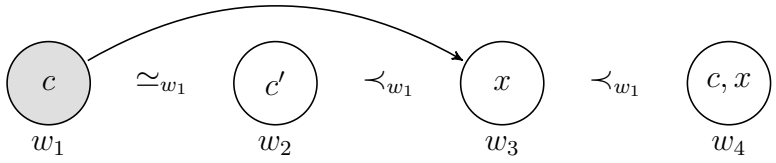

Figure 2.1: RA model for Example 1.1 (partially drawn, reflexive loops omitted)

Consider the model in Fig. 2.1, drawn for student A in Example 1.1. An arrow from w to v indicates that $w \twoheadrightarrow v$, i.e., v is uneliminated by the agent in w. (For all $v \in W$, $v \twoheadrightarrow v$, but we omit all reflexive loops.) The ordering of the worlds by their

relevance at w_1, which we take to be the actual world, is indicated between worlds.[16] In w_1, the patient has the common condition c, represented by the atomic sentence c true at w_1.[17] Possibility w_2, in which the patient has the other common condition c' instead of c, is just as relevant as w_1. Since the model is for student A, who ran the lab tests to rule out c', A has eliminated w_2 in w_1. A more remote possibility than w_2 is w_3, in which the patient has the rare disease x. Since A has not run any tests to rule out x, A has not eliminated w_3 in w_1. Finally, the most remote possibility of all is w_4, in which the patient has both c and x. We assume that A has learned from textbooks that x confers immunity to c, so A has eliminated w_4 in w_1.

Now consider C-semantics. In discussing Example 1.1, we held that student A knows that the patient's condition is c, despite the fact that A did not rule out the remote possibility of the patient's having x. C-semantics issues the opposite verdict. According to C-semantics, Kc is true at w_1 iff *all* $\neg c$-worlds, regardless of their relevance, are ruled out by the agent in w_1. However, w_3 is not ruled out by A in w_1, so Kc is false at w_1. Nonetheless, A has some knowledge in w_1. For example, one can check that $K(\neg x \rightarrow c)$ is true at w_1.

Remark 2.2 (Skepticism)**.** A skeptic might argue, however, that we have failed to include in our model a particular possibility, far-fetched but uneliminated, in which the patient has neither x nor c, the inclusion of which would make even $K(\neg x \rightarrow c)$ false at w_1 according to C-semantics. In this way, C-semantics plays into the hands of skeptics. By contrast, L- and D-semantics help to avoid skepticism by not requiring the elimination of every far-fetched possibility.

Consider the model in Fig. 2.1 from the perspective of L-semantics. According to L-semantics, student A *does* know that the patient has condition c. Kc is true at w_1, because c is true in all of the most relevant and uneliminated (at w_1) worlds, namely w_1 itself. Moreover, although A has not ruled out the possibility w_3 in which the patient has disease x, according to L-semantics she nonetheless *knows* that the

[16] We ignore the relevance orderings for the other worlds. We also ignore which possibilities are ruled out at worlds other than w_1, since we are not concerned here with student A's higher-order knowledge at w_1. If we were, then we would include other worlds in the model.

[17] Recall the double use of 'c', 'c'', and 'x' explained in footnote 3 in Chapter 1.

patient does not have x. $K\neg x$ is true at w_1, because $\neg x$ is true in all of the most relevant (at w_1) worlds: w_1 and w_2. Indeed, note that $K\neg x$ would be true at w_1 no matter how we defined the $\twoheadrightarrow$ relation.

Remark 2.3 (Vacuous Knowledge)**.** What this example shows is that according to L-semantics, in some cases an agent can know some φ *with no requirement of ruling out possibilities*, i.e., with no requirement on $\twoheadrightarrow$, simply because none of the accessible $\neg\varphi$-possibilities are relevant at w, i.e., because they are not in $\mathrm{Min}_{\preceq_w}(W)$. This is the position of Stine [1976, 257] and Rysiew [2006, 265], who hold that one can know that skeptical hypotheses do not obtain, without any evidence, simply because the skeptical possibilities are not relevant in the context (also see Lewis 1996, 561f). In general, on the kind of $\mathsf{RS}_{\exists\forall}$ view represented by L-semantics, an agent can know a *contingent empirical truth* φ with no requirement of empirically eliminating any possibilities. Heller [1999a, 207] rejects such "vacuous knowledge," and in §4.1 I discuss this Problem of Vacuous Knowledge at length (also see Cohen 1988, 99; Vogel 1999, 158f; and Remark 2.5 below). By contrast, on the kind of $\mathsf{RS}_{\forall\exists}$ view represented by D-semantics, as long as there is an accessible $\neg\varphi$-possibility, there will be some most relevant (at w) $\neg\varphi$-possibility that the agent must rule out in order to know φ in w. Hence D-semantics avoids vacuous knowledge.

D-semantics avoids the skepticism of C-semantics and the vacuous knowledge of L-semantics, but at a cost for closure. Consider the model in Fig. 2.1 from the perspective of D-semantics. First observe that D-semantics issues our original verdict that student A knows the patient's condition is c. Kc is true at w_1 since the most relevant (at w_1) $\neg c$-world, w_2, is ruled out by A in w_1. $K(c \to \neg x)$ is also true at w_1, since the most relevant (at w_1) $\neg(c \to \neg x)$-world, w_4, is ruled out by A in w_1. Not only that, but $K(c \leftrightarrow \neg x)$ is true at w_1, since the most relevant (at w_1) $\neg(c \leftrightarrow \neg x)$-world, w_2, is ruled out by A in w_1. However, the most relevant (at w_1) x-world, w_3, is *not* ruled out by A in w_1, so $K\neg x$ is false at w_1. Hence A does not know that the patient does not have disease x. We have just established the second part of the following fact, which matches Dretske's [1970] view. The first part, which follows directly from the truth definition, matches Lewis's [1996, 563n21] view.

Fact 2.2 (Known Implication)**.** The principles $K\varphi \wedge K(\varphi \to \psi) \to K\psi$ and $K\varphi \wedge K(\varphi \leftrightarrow \psi) \to K\psi$ are C/L-valid, but not D-valid.[18]

In Dretske's [1970, 1007] terminology, Fact 2.2 shows that the knowledge operator in D-semantics is not *fully penetrating*, since it does not penetrate to all of the logical consequence of what is known. Yet Dretske claims that the knowledge operator is *semi-penetrating*, since it does penetrate to some logical consequences: "it seems to me fairly obvious that if someone knows that P and Q, he thereby knows that Q" and "If he knows that P is the case, he knows that P or Q is the case" (1009). This is supposed to be the "trivial side" of Dretske's thesis (ibid.). However, if we understand the RA theory according to D-semantics, then even these monotonicity principles fail (as they famously do for Nozick's theory, discussed in §2.5, for the same structural reasons).

Fact 2.3 (Distribution & Addition)**.** The principles $K(\varphi \wedge \psi) \to K\varphi \wedge K\psi$ and $K\varphi \to K(\varphi \vee \psi)$ are C/L-valid, but not D-valid.

Proof. The proof of C/L-validity is routine. For D-semantics, the pointed model $\mathcal{M}, w_1$ in Fig. 2.1 falsifies $K(c \wedge \neg x) \to K\neg x$ and $Kc \to K(c \vee \neg x)$. These principle are of the form $K\alpha \to K\beta$. In both cases, the most relevant (at w_1) $\neg\alpha$-world in $\mathcal{M}$ is w_2, which is eliminated by the agent in w_1, so $K\alpha$ is true at w_1. However, in both cases, the most relevant (at w_1) $\neg\beta$-world in $\mathcal{M}$ is w_3, which is uneliminated by the agent in w_1, so $K\beta$ is false at w_1. □

Fact 2.3 is only the tip of the iceberg, the full extent of which is revealed in §2.6. But it already points to a dilemma. On the one hand, if we understand the RA theory according to D-semantics, then the knowledge operator lacks even the basic closure properties that Dretske wanted from a semi-penetrating operator, contrary to the "trivial side" of his thesis; here we have an example of what I called the Problem of Containment. On the other hand, if we understand the RA theory according to L-semantics, then the knowledge operator is a fully-penetrating operator, contrary to the

[18] It is easy to see that for D-semantics (and H/N/S-semantics in §2.5), knowledge fails to be closed not only under known material implication, but even under known *strict* implication: $K\varphi \wedge K\Box(\varphi \to \psi) \to K\psi$, with the $\Box$ in Definition 2.12 (or even the universal modality).

non-trivial side of Dretske's thesis; and we have the Problem of Vacuous Knowledge. It is difficult to escape this dilemma while retaining something like Heller's [1989, 1999a] world-ordering picture with which we started before Definition 2.2. However, Dretske's [1981] discussion of relevancy sets leaves open whether the RA theory should be developed along the lines of this world-ordering picture, so we have not foreclosed other ways around the dilemma. The search for something beyond the world-ordering picture will play a major role in later chapters.

2.5 Counterfactual Belief (CB) Models

In this section, I introduce the formalizations of Heller's [1989, 1999a] RA theory, Nozick's [1981] tracking theory, and Sosa's [1999] safety theory. Let us begin by defining another class of models, closely related to RA models.

Definition 2.6 (CB Model)**.** A *counterfactual belief model* is a tuple $\mathcal{M}$ of the form $\langle W, D, \leqslant, V\rangle$ where W, $\leqslant$, and V are defined in the same way as W, $\preceq$, and V for RA models in Definition 2.2, and D is a serial binary relation on W.

Notation 2.1 and Definition 2.3 apply to CB models as for RA models, but with $\leqslant_w$ in place of $\preceq_w$, $<_w$ in place of $\prec_w$, and $\equiv_w$ in place of $\simeq_w$.

Think of D as a *doxastic accessibility* relation, so that wDv indicates that everything the agent believes in w is true in v [Lewis, 1986, §1.4]. For convenience, let us extend the epistemic language of Definition 2.1 to an *epistemic-doxastic* language with a belief operator B for the D relation. We do so in order to state perspicuous truth definitions for the K operator, which could be equivalently stated in a more direct (though cumbersome) way in terms of the D relation. Our main result concerning closure will be given for the pure epistemic language without B.

Think of $\leqslant_w$ either as a relevance relation as in §2.4 (for Heller) or as a relation of *comparative similarity* with respect to world w, used for assessing counterfactuals as in Lewis 1973.[19] With the latter interpretation, we can capture the following

[19]Heller [1989] argues that the orderings for relevance and similarity are the same, only the boundary of the relevant worlds that one must rule out in order to know extends beyond that of the most similar worlds. See the remarks in §2.B below, which apply here as well.

well-known counterfactual conditions on an agent's belief that φ:

- if φ were false, the agent would not believe φ (*sensitivity*);
- if φ were true, the agent would believe φ (*adherence*);
- the agent would believe φ only if φ were true (*safety*).

Nozick [1981] argued that sensitivity and adherence—the conjunction of which is *tracking*—are necessary and sufficient for one's belief to constitute knowledge,[20] while Sosa [1999] argued that safety is necessary. (In §2.D, I will consider the revised sensitivity and safety conditions that take into account methods and bases of belief.) Following Nozick and Sosa, I interpret sensitivity as the counterfactual $\neg\varphi \boxrightarrow \neg B\varphi$, adherence as $\varphi \boxrightarrow B\varphi$, and safety as $B\varphi \boxrightarrow \varphi$, with the caveat in Observation 2.1 below. I will understand the truth of counterfactuals following Lewis [1973, 20],[21] such that $\varphi \boxrightarrow \psi$ is true at a world w iff the closest φ-worlds to w according to $\leqslant_w$ are ψ-worlds, subject to the same caveat.[22] As I explain in §2.B, the formalization is also compatible with the view that the conditions above should be understood in terms of "close enough" rather than closest worlds.

We are now prepared to define three more semantics for the K operator: H-semantics for **H**eller, N-semantics for **N**ozick, and S-semantics for **S**osa.

Remark 2.4 (Necessary Conditions)**.** In defining these semantics, I assume that each theory proposes necessary and sufficient conditions for knowledge. This is true of Nozick's [1981] theory, as it was of Lewis's [1996], but Sosa [1999] and Heller [1999a] propose only necessary conditions. Hence one may choose to read $K\varphi$ as "the agent *safely believes* φ/has *ruled out the relevant alternatives to* φ" for S/H-semantics. Our results for S/H-semantics can then be viewed as results about the logic of safe belief/the logic of relevant alternatives. However, for reasons similar to those given

[20] Nozick used the term 'variation' for what I call 'sensitivity' and used 'sensitivity' to cover both variation and adherence; but the narrower use of 'sensitivity' is now standard.

[21] Unlike Lewis, I assume $\leqslant_w$ is well-founded and only weakly centered on w (Definition 2.3).

[22] Nozick [1981, 680n8] tentatively proposes alternative truth conditions for counterfactuals. However, he also indicates that his theory may be understood in terms of Lewis's semantics for counterfactuals (but see Observation 2.1). This has become the standard practice in the literature. For example, see Vogel 1987, Comesaña 2007, and Alspector-Kelly 2011.

by Brueckner [2004] and Murphy [2006], I argue in §2.C that if the subjunctivist or RA conditions are treated as necessary for knowledge, then closure failures for these conditions threaten closure for knowledge itself. It is up to defenders of these theories to explain why knowledge is closed in ways their conditions on knowledge are not.

Definition 2.7 (Truth in a CB Model)**.** Given a well-founded CB model $\mathcal{M} = \langle W, D, \leqslant, V\rangle$ with $w \in W$ and φ in the epistemic-doxastic language, define $\mathcal{M}, w \vDash_x \varphi$ as follows (with propositional cases as in Definition 2.4):

$$\mathcal{M}, w \vDash_x B\varphi \quad \text{iff} \quad \forall v \in W\text{: if } wDv \text{ then } \mathcal{M}, v \vDash_x \varphi;$$

$$\begin{aligned}\mathcal{M}, w \vDash_h K\varphi \quad \text{iff} \quad & \mathcal{M}, w \vDash_h B\varphi \text{ and} \\ & \text{(sensitivity) } \forall v \in \mathrm{Min}_{\leqslant_w}\big(\overline{[\![\varphi]\!]}_h\big)\text{: } \mathcal{M}, v \nvDash_h B\varphi;\end{aligned}$$

$$\begin{aligned}\mathcal{M}, w \vDash_n K\varphi \quad \text{iff} \quad & \mathcal{M}, w \vDash_n B\varphi \text{ and} \\ & \text{(sensitivity) } \forall v \in \mathrm{Min}_{\leqslant_w}\big(\overline{[\![\varphi]\!]}_n\big)\text{: } \mathcal{M}, v \nvDash_n B\varphi, \\ & \text{(adherence) } \forall v \in \mathrm{Min}_{\leqslant_w}\big([\![\varphi]\!]_n\big)\text{: } \mathcal{M}, v \vDash_n B\varphi;\end{aligned}$$

$$\begin{aligned}\mathcal{M}, w \vDash_s K\varphi \quad \text{iff} \quad & \mathcal{M}, w \vDash_s B\varphi \text{ and} \\ & \text{(safety) } \forall v \in \mathrm{Min}_{\leqslant_w}\big([\![B\varphi]\!]_s\big)\text{: } \mathcal{M}, v \vDash_s \varphi.\end{aligned}$$

Note that the truth clause for $B\varphi$ guarantees *doxastic* closure (see Fact 2.7).[23]

It is easy to check that the belief and subjunctive conditions of H/N/S-semantics together ensure Fact 2.4 (cf. Heller 1999b, 126; Kripke 2011, 164).

[23] It is not essential that we model belief with a doxastic accessibility relation. When we show that a given closure principle is H/N/S-*valid*, we use the fact that the truth clause for $B\varphi$ in Definition 2.7 guarantees some *doxastic closure* (see Fact 2.7); but when we show that a closure principle is *not* H/N/S-valid, we do not use any facts about doxastic closure, as one can verify by inspection of the proofs. For the purpose of demonstrating closure failures, we could simply associate with each $w \in W$ a set of formulas Σ_w such that $\mathcal{M}, w \vDash B\varphi$ iff $\varphi \in \Sigma_w$. However, if we were to assume no doxastic closure properties for Σ_w, then there would be no valid epistemic closure principles (except $K\varphi \to K\varphi$), assuming knowledge requires belief. As a modeling choice, this may be realistic, but it throws away information about the reasons for closure failures. For we would no longer be able to tell whether an epistemic closure principle such as $K\varphi \to K(\varphi \vee \psi)$ is not valid for the (interesting) reason that the special conditions for knowledge posited by a theory are not preserved in the required way, or whether the principle is not valid for the (uninteresting) reason that there is some agent who knows φ but happened not to form a belief in $\varphi \vee \psi$.

Fact 2.4 (Veridicality)**.** $K\varphi \to \varphi$ is H/N/S-valid.

Observation 2.1 (Adherence and Safety)**.** The adherence condition in the N-semantics clause may be equivalently replaced by

$$\forall v \in \mathrm{Min}_{\leqslant_w}(W)\colon \mathcal{M}, v \vDash_n \varphi \to B\varphi;$$

the safety condition in the S-semantics clause may be equivalently replaced by

$$\forall v \in \mathrm{Min}_{\leqslant_w}(W)\colon \mathcal{M}, v \vDash_s B\varphi \to \varphi.$$

This observation has two important consequences. The first is that in *centered* models (Def. 2.3.5), adherence ($\varphi \mathrel{\Box\!\!\to} B\varphi$) and safety ($B\varphi \mathrel{\Box\!\!\to} \varphi$) add nothing to belief and true belief, respectively, given standard Lewisian semantics for counterfactuals. DeRose [1995, 27n27] takes adherence to be redundant apparently for this reason. But since we only assume *weak* centering, adherence as above makes a difference—obviously for truth in a model, but also for validity (see Fact 2.9). Nozick [1981, 680n8] suggests another way of understanding adherence so that it is non-trivial, but here I will settle on its simple interpretation with weak centering in standard semantics. Whether or not weak centering is right for counterfactuals, adherence and safety can be—and safety typically is—understood directly in terms of what holds in a set of close worlds including the actual world, our $\mathrm{Min}_{\leqslant_w}(W)$ (see note 2.B), rather than as $\varphi \mathrel{\Box\!\!\to} B\varphi$ and $B\varphi \mathrel{\Box\!\!\to} \varphi$.[24] (Adherence is typically ignored.) For sensitivity alone, centering vs. weak centering makes no difference for valid principles.

The second consequence is that safety is a $\exists\forall$ condition as in §2.4, where $\mathrm{Min}_{\leqslant_w}(W)$

[24] Alternatively, the sphere of worlds for adherence could be independent of the relation $\leqslant_w$ for sensitivity, i.e., distinct from $\mathrm{Min}_{\leqslant_w}(W)$ (see Remark 3.2), so $\leqslant_w$ could be centered without trivializing adherence. But this would allow cases in which an agent knows φ even though she believes φ in a $\neg\varphi$-world that is "close enough" to w to be in its adherence sphere (provided there is a closer $\neg\varphi$-world according to $\leqslant_w$ in which she does not believe φ). Nozick [1981, 680n8] suggests interpreting adherence counterfactuals $\varphi \mathrel{\Box\!\!\to} B\varphi$ with true antecedents in such a way that the sphere over which $\varphi \to B\varphi$ must hold may differ for different φ. By contrast, Observation 2.1 shows that we are interpreting adherence as a kind of $\exists\forall$ condition in a sense that generalizes that of §2.4 to cover a requirement that one meet an epistemic success condition in all P-worlds in $\mathsf{R}_c(w)$ (see §3.3.2). A $\forall\exists$ interpretation of adherence that, e.g., allows the adherence sphere for $\varphi \vee \psi$ to go beyond that of φ, would create another source of closure failure (see §2.6.5 and §2.10).

serves as the set $\mathsf{R}_c(w)$ that is independent of the particular proposition in question (cf. Alspector-Kelly 2011, 129n6). By contrast, sensitivity is obviously a $\forall\exists$ condition, analogous to the D-semantics clause. Viewed this way, in the "subjunctivist-flavored" family of D/H/N/S-semantics, S-semantics is the odd member of the family, since by only looking at the fixed set $\mathrm{Min}_{\leqslant_w}(W)$ in the safety clause, it never uses the rest of the world-ordering.[25]

Fig. 2.2 displays a CB model for Example 1.1. The dotted arrows represent the doxastic relation D. That the only arrow from w_1 goes to itself indicates that in w_1, student A believes that the actual world is w_1, where the patient has c and not x. (We do not require that D be functional, but in Fig. 2.2 it is.) Hence $\mathcal{M}, w_1 \vDash B(c \wedge \neg x)$. That the only arrow from w_3 goes to w_1 indicates that in w_3, A believes that w_1 is the actual world; since w_3 is the closest (to w_1) x-world, we take this to mean that if the patient's condition were x, A would still believe it was c and not x (because A did not run any of the tests necessary to detect x).[26] Hence $\mathcal{M}, w_1 \nvDash_{h,n} K\neg x$, because the *sensitivity* condition is violated. However, one can check that $\mathcal{M}, w_1 \vDash_{h,n} Kc$.

If we were to draw the model for student B, we would replace the arrow from w_2 to w_2 by one from w_2 to w_1, reflecting that if the patient's condition were c', B would still believe it was c (because B made the diagnosis of c after only a physical exam, and c and c' have the same visible symptoms). Hence $\mathcal{M}', w_1 \nvDash_{h,n} Kc$, where $\mathcal{M}'$ is the model with $w_2 D w_1$ instead of $w_2 D w_2$.

When we consider S-semantics, we get a different verdict on whether A knows that the patient does not have disease x. Observe that $\mathcal{M}, w_1 \vDash_s K\neg x$, because student A believes $\neg x$ in w_1 and at the closest worlds to w_1, namely w_1 and w_2, $\neg x$ is true. Therefore, A safely believes $\neg x$ in w_1. Similarly $\mathcal{M}, w_1 \vDash_s Kc$, because A safely

[25] Note that safety and tracking theorists may draw different models, with different $\leqslant_w$ relations and $\mathrm{Min}_{\leqslant_w}(W)$ sets, to represent the epistemic situation of the same agent.

[26] What about w_4? In §2.4, we assumed that A learned from textbooks that x confers immunity to c, so she had eliminated w_4 at w_1. In Fig. 2.2, that the only arrow from w_4 goes to w_4 indicates that if (contrary to biological law) x did *not* confer immunity to c and the patient had both c and x, then A would believe that he had both c and x, perhaps because the textbooks and tests would be different in such a world. However, all we need to assume for the example to work is that if the patient had both c and x, then it would be *compatible* with what A believes that the patient had both c and x, reflected by the reflexive loop. We can have other outgoing arrows from w_4 as well.

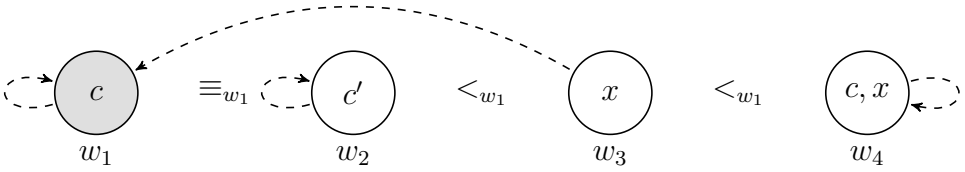

Figure 2.2: CB model for Example 1.1 (partially drawn)

believes c in w_1. Yet if we add the arrow from w_2 to w_1 for B, one can check that B does not safely believe c at w_1, so $\mathcal{M}', w_1 \not\models_s Kc$.

Remark 2.5 (Vacuous Knowledge Again)**.** The fact that $\mathcal{M}, w_1 \models_s K\neg x$ reflects the idea that the safety theory leads to a *neo-Moorean* response to skepticism [Sosa, 1999], according to which agents can know that skeptical possibilities do not obtain. In general, a point parallel to that of Remark 2.3 holds for the $\mathsf{RS}_{\exists\forall}$ safety theory: if the $\neg\varphi$-worlds are not among the close worlds, then one's belief in φ is automatically safe, no matter how poorly one's beliefs match the facts in possible worlds (cf. Alspector-Kelly's [2011] distinction between near-safe and far-safe beliefs). This is the version of the Problem of Vacuous Knowledge for the safety theory (see §4.1.4). By contrast, on the kind of $\mathsf{RS}_{\forall\exists}$ theory represented by H/N-semantics, if $\neg\varphi$ is possible, then knowledge requires that one not falsely believe φ in the closest $\neg\varphi$-worlds.

Like D-semantics, H/N-semantics avoid the skepticism of C-semantics and the vacuous knowledge of L/S-semantics, but at a cost for closure. All of the closure principles shown in Facts 2.2 and 2.3 to be falsifiable in RA models under D-semantics are also falsifiable in CB models under H/N-semantics, as one can check at w_1 in Fig. 2.2. After embracing the "nonclosure" of knowledge under known implication, Nozick [1981, 231ff] tried to distinguish successful from unsuccessful cases of knowledge transmission by whether extra subjunctive conditions hold;[27] but doing so does

[27]Roughly, Nozick [1981, 231ff] proposes than an agent knows ψ via inference from φ iff (1) $K\varphi$, (2) she infers the true conclusion ψ from premise φ, (3) $\neg\psi \boxdot\!\!\rightarrow \neg B\varphi$, and (4) $\psi \boxdot\!\!\rightarrow B\varphi$. Whether this proposal is consistent with the rest of Nozick's theory depends on whether (1) - (4) ensure that the agent tracks ψ, which is still necessary for her to know ψ (234); and that depends on what kind of modal connection between $B\varphi$ and $B\psi$ is supposed to follow from (2), because (1), (3), and (4) together do not ensure that she tracks ψ.

not eliminate the unsuccessful cases, which go far beyond nonclosure under known implication, as shown in §2.6.

Nozick was well aware that $K(\varphi \wedge \psi) \rightarrow K\varphi \wedge K\psi$ fails on his theory, and his explanation (beginning "S's belief that $p\&q$ ..." on 228) is similar to a proof in our framework. He resisted the idea that $K\varphi \rightarrow K(\varphi \vee \psi)$ fails, but he is clearly committed to it.[28] Vogel's [2007, 76] explanation of why it fails for Nozick is also similar to a proof in our framework, as are Kripke's [2011] many demonstrations of closure failure for Nozick's theory. Partly in response to these problems, Roush [2005, 2012] proposes a *recursive tracking* view of knowledge, in a probabilistic framework, with an additional recursion clause to support closure (see §2.C). For discussion of the relation between probabilistic and subjunctivist versions of tracking, see §2.E.

All of the closure principles noted fail for S-semantics as well. For example, it is easy to construct a model in which $B(\varphi \wedge \psi)$ and hence $B\varphi$ are true at a world w, all worlds close to w satisfy $B(\varphi \wedge \psi) \rightarrow \varphi \wedge \psi$, and yet some worlds close to w do not satisfy $B\varphi \rightarrow \varphi$, resulting in a failure of $K(\varphi \wedge \psi) \rightarrow K\varphi$ at w. Murphy's [2005, 2006, §4.3] intuitive examples of closure failure for safety have exactly this structure.[29] We return to this problem for safety in §2.10.

Now it is time to go beyond case-by-case assessment of closure principles. In the following sections, we will turn to results of a more general nature.

2.6 The Closure Theorem and Its Consequences

In this section, I state the main result of the chapter, Theorem 2.1, which characterizes the closure properties of knowledge for the theories we have formalized. Despite

[28] While Nozick [1981] admits that such a closure failure "surely carries things too far" (230n64, 692), he also says that an agent can know p and yet fail to know $\neg(\neg p \wedge \mathrm{SK})$ (228). But the latter is logically equivalent to $p \vee \neg\mathrm{SK}$, and Nozick accepts closure under (known) logical equivalence (229). Nozick suggests (236) that closure under deducing a disjunction from a disjunct should hold, provided methods of belief formation are taken into account. However, §2.D shows that taking methods into account does not help here.

[29] For Murphy's [2006, §4.3] "Lying Larry" example, take φ to be *Larry is married* and ψ to be *Larry is married to Pat*. For Murphy's [2005, 333] variation on Kripke's red barn example, take φ to be *the structure is a barn* and ψ to be *the structure is red*.

the differences between the RA, tracking, and safety theories of knowledge as formalized by D/H/N/S-semantics, Theorem 2.1 provides a unifying perspective: the valid epistemic closure principles are essentially the same for these different theories, except for a twist with the theory of *total* RA models. For comparison, I also include C/L-semantics, which fully support closure.

Formally, Theorem 2.1 is the same type of result as the "modal decomposition" results of van Benthem [2010, §4.3, §10.4] for the weakest normal modal logic **K** and the weakest monotonic modal logic **M** (see Chellas 1980, §8.2). From Theorem 2.1 we obtain decidability (Corollary 2.1) and completeness (Corollary 2.4) results as corollaries. From the proof of the theorem, we obtain results on finite models (Corollary 2.2) and complexity (Corollary 2.3).

The following notation will be convenient throughout this section.

Notation 2.3 (Closure Notation)**.** Given (possibly empty) sequences of formulas $\varphi_1, \ldots, \varphi_n$ and $\psi_1, \ldots, \psi_m$ in the epistemic language and a propositional conjunction φ_0, we use the notation

$$\chi_{n,m} := \varphi_0 \wedge K\varphi_1 \wedge \cdots \wedge K\varphi_n \to K\psi_1 \vee \cdots \vee K\psi_m.$$

Call such a $\chi_{n,m}$ a *closure principle*.[30]

Hence a closure principle states that if the agent knows each of φ_1 through φ_n (and the world satisfies a non-epistemic φ_0), then the agent knows at least one of ψ_1 through ψ_m. Our question is: which closure principles are *valid*?

Theorem 2.1 is the answer. Its statement refers to a "T-unpacked" closure principle, a notion I have not yet introduced. For the first reading of the theorem, this can be ignored. Think only of *flat* formulas without nesting of the K operator (Def. 2.1), which are T-unpacked if $\varphi_1 \wedge \cdots \wedge \varphi_n$ is a conjunct of φ_0. Or we can ignore T-unpacking for flat $\chi_{n,m}$ and replace condition (a) of the theorem by

(a)′ $\varphi_0 \wedge \cdots \wedge \varphi_n \to \bot$ is valid.

[30]Following standard convention, we take an empty disjunction to be $\bot$, so a closure principle $\chi_{n,0}$ with no $K\psi$ formulas is of the form $\varphi_0 \wedge K\varphi_1 \wedge \cdots \wedge K\varphi_n \to \bot$.

Example 2.1 will show the need for T-unpacking, defined in general in §2.6.2.

Theorem 2.1 (Closure Theorem)**.** Let

$$\chi_{n,m} := \varphi_0 \wedge K\varphi_1 \wedge \cdots \wedge K\varphi_n \to K\psi_1 \vee \cdots \vee K\psi_m$$

be a T-unpacked closure principle.

1. $\chi_{n,m}$ is C/L-valid over relevant alternatives models iff

   (a) $\varphi_0 \to \bot$ is valid or

   (b) for some $\psi \in \{\psi_1, \ldots, \psi_m\}$,

   $$\varphi_1 \wedge \cdots \wedge \varphi_n \to \psi \text{ is valid;}$$

2. $\chi_{n,m}$ is D-valid over total relevant alternatives models iff (a) or

   (c) for some $\Phi \subseteq \{\varphi_1, \ldots, \varphi_n\}$ and nonempty $\Psi \subseteq \{\psi_1, \ldots, \psi_m\}$,[31]

   $$\bigwedge_{\varphi \in \Phi} \varphi \leftrightarrow \bigwedge_{\psi \in \Psi} \psi \text{ is valid;}$$

3. $\chi_{n,m}$ is D-valid over all relevant alternatives models iff (a) or

   (d) for some $\Phi \subseteq \{\varphi_1, \ldots, \varphi_n\}$ and $\psi \in \{\psi_1, \ldots, \psi_m\}$,

   $$\bigwedge_{\varphi \in \Phi} \varphi \leftrightarrow \psi \text{ is valid;}$$

4. $\chi_{n,m}$ is H/N/S-valid over counterfactual belief models if (a) or (d);[32] and a flat $\chi_{n,m}$ is H/N/S-valid over such models only if (a) or (d).

[31] Following standard convention, if $\Phi = \emptyset$, we take $\bigwedge_{\varphi \in \Phi} \varphi$ to be $\top$.

[32] When I refer to (d) from part 4, I mean the condition that $\bigwedge_{\varphi \in \Phi} \varphi \leftrightarrow \psi$ is H/N/S-valid.

The remarkable fact established by Theorem 2.1 that D/H/N/S-semantics validate essentially the same closure principles, except for the twist of totality in (c), further supports talk of their representing a "family" of subjunctivist-flavored theories of knowledge. Although results in §2.9.2 (Facts 2.9.4, 2.9.5, and 2.11.1) show that the 'only if' direction of part 4 does not hold for some principles involving higher-order knowledge, the agreement between D/H/N/S-semantics on the validity of flat closure principles is striking.

Remark 2.6 (Independence from Assumptions)**.** Recalling the types of orderings in Definition 2.3, it is noteworthy that parts 1 and 4 of Theorem 2.1 are independent of whether we assume totality (or universality), while parts 2 and 3 are independent of whether we assume centering, linearity (see §2.6.2), or universality (see Prop. 2.3). For parts 1 - 4, we can drop our running assumption of well-foundedness, provided we modify the truth definitions accordingly (see Remark 2.7). Finally, part 1 for L-semantics (but not C-semantics) and parts 2 - 3 for D-semantics are independent of additional properties of $\rightarrow\!\!\!\!\circ$ such as transitivity and symmetry (see Remark 2.8 and Example 2.3).

To apply the theorem, observe that $Kp \wedge K(p \rightarrow q) \rightarrow Kq$ is not D/H/N/S-valid, because $p \wedge (p \rightarrow q) \rightarrow \bot$ is not valid, so (a)$'$ fails, and none of

$$p \wedge (p \rightarrow q) \leftrightarrow q,\ p \leftrightarrow q,\ (p \rightarrow q) \leftrightarrow q,\ \text{or } \top \leftrightarrow q$$

are valid, so there are no Φ and Ψ/ψ as described. Hence (c)/(d) fails.

On the other hand, we now see that $Kp \wedge Kq \rightarrow K(p \wedge q)$ *is* D/H/N/S-valid, because $p \wedge q \leftrightarrow p \wedge q$ is valid, so we can take $\Phi = \{p, q\}$ and $\Psi = \{p \wedge q\}$ or $\psi = p \wedge q$. Besides $K\varphi \rightarrow \varphi$ (Facts 2.1 and 2.4), this is the first *valid* principle we have identified for D/H/N/S-semantics, to which we will return in §2.8.

Fact 2.5 (C Axiom)**.** The principle $K\varphi \wedge K\psi \rightarrow K(\varphi \wedge \psi)$, known as the C axiom, is D/H/N/S-valid.

To get a feel for Theorem 2.1, it helps to test a variety of closure principles.

Exercise 2.1 (Testing Closure). Using Theorem 2.1, verify that neither $K(p \wedge q) \rightarrow K(p \vee q)$ nor $Kp \wedge Kq \rightarrow K(p \vee q)$ are D/H/N/S-valid; verify that $K(p \wedge q) \rightarrow Kp \vee Kq$ is only D-valid over *total* RA models; verify that $K(p \vee q) \wedge K(p \rightarrow q) \rightarrow Kq$ and $Kp \wedge K(p \rightarrow q) \rightarrow K(p \wedge q)$ are D/H/N/S-valid.

As if the closure failures of Fact 2.3 were not bad enough, the first three of Exercise 2.1 are also highly counterintuitive. Recall from §2.1 that the Dretske-Nozick case against full closure under known implication, K, had two parts: examples in which K purportedly fails, such as Example 1.1, and theories of knowledge that purportedly explain the failures. For the other principles, we can see why they fail according to the subjunctivist-flavored theories; but without some intuitive examples in which, e.g., arguably an ideally astute logician knows two propositions but not their disjunction, the failure of such weak closure principles according to a theory of knowledge seems to be strong evidence against the theory—even for those sympathetic to the denial of K.

While the closure failures permitted by subjunctivist-flavored theories go too far, in another way they do not go far enough for some purposes. Reflection on the last two principles of Exercise 2.1 suggests they are about as dangerous as K in arguments for radical skepticism about knowledge. The fact that one's theory validates these principles seems to undermine the force of one's denying K in response to skepticism, as Nozick [1981] uses his subjunctivism to do.

Notwithstanding these negative points against subjunctivist-flavored theories of *knowledge*, simply replace the K symbol in our language by a neutral $\Box$ and Theorem 2.1 can be viewed as a neutral result about the logic of relevant alternatives, of sensitive/truth-tracking belief, and of safe belief (see §2.8).

Parts 3 and 4 of Theorem 2.1 reflect that D-semantics over RA models and H/N/S-semantics over CB models have the following *separation property*.

Proposition 2.1 (Separation)**.** For D-semantics (resp. H/N/S-semantics), a closure principle $\chi_{n,m}$ (resp. a flat $\chi_{n,m}$) as in Notation 2.3 with $m \geq 1$ is valid iff there is some $j \leq m$ such that $\varphi_0 \wedge K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi_j$ is valid.

The reason for this separation property comes out clearly in the proofs in §2.6.3

and §2.6.4. In essence, if the principles with single disjunct consequents are all invalid, then we can glue their falsifying models together to obtain a falsifying model for $\chi_{n,m}$. However, this is not the case for D-semantics over *total* RA models. The following fact demonstrates the nonequivalence of D-semantics over total RA models and D-semantics over all RA models (as well as H/N/S-semantics over total/all CB models) with an interesting new axiom.

Fact 2.6 (X Axiom)**.** The principle $K(\varphi \wedge \psi) \rightarrow K\varphi \vee K\psi$, hereafter called the "X axiom" (see §2.8), is D-valid over total RA models, but not D-valid over all RA models or H/N/S-valid over (total) CB models.

Proof. I leave D-validity over total RA models to the reader. Fig. 2.3 displays a non-total RA model that falsifies $K(p \wedge q) \rightarrow Kp \vee Kq$ in D-semantics. Since $\mathrm{Min}_{\preceq_w}(\overline{[\![p \wedge q]\!]}) = \{v, x\}$, $w \not\nrightarrow v$, and $w \not\nrightarrow x$, $\mathcal{M}, w \vDash_d K(p \wedge q)$. Since u and x are incomparable according to $\preceq_w$, as are y and v, we have $u \in \mathrm{Min}_{\preceq_w}(\overline{[\![p]\!]})$ and $y \in \mathrm{Min}_{\preceq_w}(\overline{[\![q]\!]})$, which with $w \nrightarrow u$ and $w \nrightarrow y$ implies $\mathcal{M}, w \nvDash_d Kp \vee Kq$. The counterexample for H/N/S-semantics is in Fig. 2.10, discussed in §2.10. □

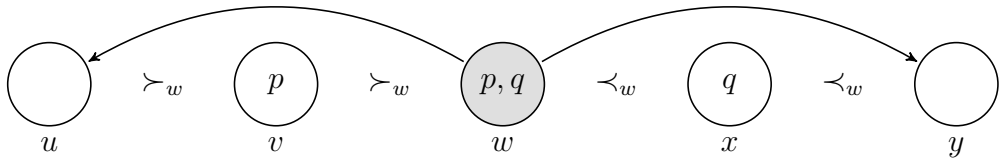

Figure 2.3: non-total RA countermodel for $K(p \wedge q) \rightarrow Kp \vee Kq$ in D-semantics (partially drawn, reflexive loops omitted)

In §2.8, we will see the role that the X axiom plays in a complete deductive system for D-semantics over total RA models, as well as the role that the C axiom plays in complete deductive systems for D/H/N/S-semantics.

Given the separation property, the proof of the 'only if' direction of Theorem 2.1.3 for *flat* closure principles can be explained roughly as follows.

Proof sketch. Let us try to falsify a flat $\varphi_0 \wedge K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi_j$. Construct a pointed model $\mathcal{M}, w$ with a valuation such that the propositional part φ_0 is true

at w.[33] To make $K\psi_j$ false while keeping all $K\varphi_i$ true at w, we want to add an *uneliminated* $\neg\psi_j$-world v such that (A) there is no $\neg\psi_j$-world more relevant than v and (B) for any $\neg\varphi_i$ true at v, there is a *more relevant* $\neg\varphi_i$-world that is *eliminated* at w. This is possible if there is a propositional valuation such that $\neg\psi_j$ is true at v and for all $\neg\varphi_i$ true at v, $\psi_j \wedge \neg\varphi_i$ is satisfiable; for then we can add a satisfying world for each conjunction and make them eliminated and more relevant than v, which gives (A) and (B). If there is no such valuation, then every valuation that satisfies $\neg\psi_j$ also satisfies some $\neg\varphi_i$ for which $\psi_j \to \varphi_i$ is valid. Then where Φ is the set of all such φ_i, $\neg\psi_j \to \bigvee_{\varphi\in\Phi} \neg\varphi$ and $\psi_j \to \bigwedge_{\varphi\in\Phi} \varphi$ are valid, which means $\bigwedge_{\varphi\in\Phi} \varphi \leftrightarrow \psi_j$ is valid. □

In §2.6.2 - 2.6.3 we give a more precise and general form of the above argument. We conclude this subsection with an example of why Theorem 2.1 requires the notion of T-unpacking, which is defined in general in Definition 2.9.

Example 2.1 (T-unpacking)**.** As noted before Theorem 2.1, if we consider only flat formulas, then we can ignore T-unpacking, provided we replace condition (a) of Theorem 2.1 by the condition: (a)′ $\varphi_1 \wedge \cdots \wedge \varphi_n \to \bot$ is valid. Let us see why T-unpacking is necessary for non-flat formulas. For example, the formula

$$KKp \wedge KKq \to K(p \wedge q) \tag{2.1}$$

is D/H/N/S-valid. Yet none of the following are valid: $Kp \wedge Kq \to \bot$, $Kp \wedge Kq \leftrightarrow p \wedge q$, $Kp \leftrightarrow p \wedge q$, $Kq \leftrightarrow p \wedge q$, and $\top \leftrightarrow p \wedge q$. Hence (2.1) does not satisfy (a)′, (c), or (d) in Theorem 2.1. However, if we *T-unpack* (2.1) by repeatedly applying the T axiom, $K\varphi \to \varphi$, to the antecedent, we obtain

$$(p \wedge q \wedge Kp \wedge Kq \wedge KKp \wedge KKq) \to K(p \wedge q), \tag{2.2}$$

which satisfies (b), (c), and (d) with $\Phi = \{p, q\}$ and $\Psi = \{p \wedge q\}$ or $\psi = p \wedge q$. Hence (2.2) is valid according to Theorem 2.1. Given the validity of the T axiom over RA/CB models (Facts 2.1 and 2.4), (2.1) and (2.2) are equivalent, so (2.1) is valid

[33] In the following argument, 'relevant' means relevant *at* w (i.e., according to $\preceq_w$) and 'uneliminated'/'eliminated' means uneliminated/eliminated *at* w (i.e., $w \twoheadrightarrow v$ or $w \not\twoheadrightarrow v$).

as well. This example shows the essential idea of T-unpacking, defined formally in §2.6.2 and demonstrate again in Example 2.2.

As shown by Proposition 2.2 below, any epistemic formula can be effectively transformed into an equivalent conjunction, each conjunct of which is a T-unpacked formula $\chi_{n,m}$ as in Notation 2.3. Using Theorem 2.1, the validity of each conjunct can be reduced to the validity of finitely many formulas of lesser modal depth (Def. 2.1). By repeating this process, we eventually obtain a finite set of propositional formulas, whose validity we can decide by truth tables. Thus, Theorem 2.1 yields the following decidability results.

Corollary 2.1 (Decidability)**.** The problem of checking whether an arbitrary formula is C/L/D-valid or whether a flat formula is H/N/S-valid over (total or all) RA/CB models is decidable.

In addition, Theorem 2.1 will yield axiomatization results in Corollary 2.4. As Corollary 2.4 will show, the 'if' direction of each 'iff' statement in Theorem 2.1 is a soundness result, while the 'only if' direction is a completeness result. We prove soundness in §2.6.1 and completeness in §2.6.2 - 2.6.4.

2.6.1 Soundness

In the 'if' direction, part 1 of Theorem 2.1 is a simple application of the C/L-truth definitions, which we skip. For parts 2 - 4, we use the following lemma.

Lemma 2.1 (Min Inclusion)**.**

1. If condition (c) of Theorem 2.1 holds, then for any well-founded and total pointed RA/CB model $\mathcal{M}, w$,[34] there is some $\psi \in \Psi$ such that

$$\mathrm{Min}_{\leq_w}(\overline{[\![\psi]\!]}) \subseteq \bigcup_{\varphi\in\Phi} \mathrm{Min}_{\leq_w}(\overline{[\![\varphi]\!]}).$$

[34]When dealing with both RA and CB models, I use $\leq_w$ to stand for $\preceq_w$ or $\leqslant_w$.

2. If condition (d) of Theorem 2.1 holds, then for any well-founded pointed RA/CB model $\mathcal{M}, w$,

$$\mathrm{Min}_{\leq_w}(\overline{[\![\psi]\!]}) \subseteq \bigcup_{\varphi\in\Phi} \mathrm{Min}_{\leq_w}(\overline{[\![\varphi]\!]}).$$

Proof. For part 1, assume for reductio that (c) holds and there is some well-founded and total $\mathcal{M}, w$ such that for all $\psi \in \Psi$ there is some u_ψ with

$$u_\psi \in \mathrm{Min}_{\leq_w}(\overline{[\![\psi]\!]}) \tag{2.3}$$

and

$$u_\psi \notin \bigcup_{\varphi\in\Phi} \mathrm{Min}_{\leq_w}(\overline{[\![\varphi]\!]}). \tag{2.4}$$

Given (c), (2.3) implies $u_\psi \in \overline{[\![\varphi_\psi]\!]}$ for some $\varphi_\psi \in \Phi$. Since $\leq_w$ is well-founded, there is some

$$v \in \mathrm{Min}_{\leq_w}(\bigcup_{\varphi\in\Phi} \overline{[\![\varphi]\!]}). \tag{2.5}$$

Given (c), (2.5) implies $v \in \overline{[\![\psi]\!]}$ for some $\psi \in \Psi$. Hence $u_\psi \leq_w v$ by (2.3) and the totality of $\leq_w$. Together $u_\psi \leq_w v$, $u_\psi \in \overline{[\![\varphi_\psi]\!]}$, (2.5), and the transitivity of $\leq_w$ imply

$$u_\psi \in \mathrm{Min}_{\leq_w}(\bigcup_{\varphi\in\Phi} \overline{[\![\varphi]\!]}), \tag{2.6}$$

which contradicts (2.4) by basic set theory.

For part 2, assume for reductio that (d) holds and there is some well-founded $\mathcal{M}, w$ and u_ψ such that (2.3) and (2.4) hold for ψ. Given (d), (2.3) implies $u_\psi \in \overline{[\![\varphi_\psi]\!]}$ for some $\varphi_\psi \in \Phi$. Hence by the well-foundedness of $\leq_w$ and (2.4) there is some $v \in \overline{[\![\varphi_\psi]\!]}$ such that $v <_w u_\psi$. Given (d), $v \in \overline{[\![\varphi_\psi]\!]}$ implies $v \in \overline{[\![\psi]\!]}$, which with $v <_w u_\psi$ contradicts (2.3). □

For the H/N/S-semantics cases, we will also use a basic fact of normal modal logic (see Theorem 3.3(2) of Chellas 1980), namely that the truth clause for B in Definition 2.7 guarantees Fact 2.7 below. Note that we do not require full doxastic closure, but only as much doxastic closure as needed to support the limited forms of epistemic

closure that are valid for H/N/S-semantics.

Fact 2.7 (Partial Doxastic Closure)**.** For $x \in \{h, n, s\}$, if $\vDash_x \bigwedge_{\varphi\in\Phi} \varphi \leftrightarrow \psi$, then $\vDash_x \bigwedge_{\varphi\in\Phi} B\varphi \leftrightarrow B\psi$.

For convenience, we will use the following notation throughout this section.

Notation 2.4 (Relational Image)**.** Given $\mathcal{M} = \langle W, \rightarrowtriangle, \preceq, V\rangle$, the image of $\{w\}$ under the relation $\rightarrowtriangle$ is $\rightarrowtriangle(w) = \{v \in W \mid w \rightarrowtriangle v\}$.

Hence $\rightarrowtriangle(w)$ is the set of uneliminated possibilities for the agent in w.

We are now ready to prove the 'if' directions of Theorem 2.1.2-4.

Claim 2.1. If (a) or (c) holds, then $\chi_{n,m}$ is D-valid over total RA models; if (a) or (d) holds, then it is D-valid over RA models and H/N/S-valid over CB models.

Proof. If (a) holds, then it is immediate that $\chi_{n,m}$ is D/H/N/S-valid, since its antecedent is always false. For (c) and (d), we consider each of the D/H/N/S-semantics in turn, assuming for an arbitrary pointed RA/CB model $\mathcal{M}, w$ that

$$\mathcal{M}, w \vDash_x \bigwedge_{\varphi\in\Phi} K\varphi. \tag{2.7}$$

To show $\mathcal{M}, w \vDash_x \chi_{n,m}$, it suffices to show $\mathcal{M}, w \vDash_x K\psi_j$ for some $j \leq m$.

If (2.7) holds for $x := d$, then by the truth definition (Def. 2.5),

$$\bigcup_{\varphi\in\Phi} \mathrm{Min}_{\preceq_w}(\overline{[\![\varphi]\!]}) \cap \rightarrowtriangle(w) = \emptyset. \tag{2.8}$$

If $\mathcal{M}$ is a total (resp. any) RA model, then by (c) and Lemma 2.1.1 (resp. by (d) and Lemma 2.1.2), (2.8) implies that there is some $\psi \in \Psi$ (resp. that the ψ in (d) is) such that $\mathrm{Min}_{\preceq_w}(\overline{[\![\psi]\!]}) \cap \rightarrowtriangle(w) = \emptyset$, whence $\mathcal{M}, w \vDash_d K\psi$.

For the cases of H/N/S-semantics, it follows from (d) and Fact 2.7 that

$$\bigcap_{\varphi\in\Phi} [\![B\varphi]\!] = [\![B\psi]\!] \text{ and } \bigcup_{\varphi\in\Phi} \overline{[\![B\varphi]\!]} = \overline{[\![B\psi]\!]}. \tag{2.9}$$

If (2.7) holds for $x := h$, then by the truth definition (Def. 2.7),

$$\mathcal{M}, w \vDash_h \bigwedge_{\varphi\in\Phi} B\varphi \text{ and } \bigcup_{\varphi\in\Phi} \mathrm{Min}_{\leqslant_w}(\overline{[\![\varphi]\!]}) \subseteq \bigcup_{\varphi\in\Phi} \overline{[\![B\varphi]\!]}. \tag{2.10}$$

By (2.9), the first conjunct of (2.10) implies $\mathcal{M}, w \vDash_h B\psi$. By (d), Lemma 2.1.2, and (2.9), the second conjunct implies the sensitivity condition that $\mathrm{Min}_{\leqslant_w}(\overline{[\![\psi]\!]}) \subseteq \overline{[\![B\psi]\!]}$. Hence $\mathcal{M}, w \vDash_h K\psi$.

If (2.7) holds for $x := n$, then by the truth definition (Def. 2.7), (2.10) holds with n in place of h. So by the same argument as before, sensitivity holds for ψ at w, which with $\mathcal{M}, w \vDash_n B\psi$ and $w \in \mathrm{Min}_{\leqslant_w}(W)$ (Def. 2.2.3b) implies $\mathcal{M}, w \vDash_n \psi$. It follows that $\mathrm{Min}_{\leqslant_w}([\![\psi]\!]) \subseteq \mathrm{Min}_{\leqslant_w}(W)$, which with (d) implies

$$\mathrm{Min}_{\leqslant_w}([\![\psi]\!]) \subseteq \bigcap_{\varphi\in\Phi} \mathrm{Min}_{\leqslant_w}([\![\varphi]\!]). \tag{2.11}$$

Since the adherence condition must hold for each $\varphi \in \Phi$ at w,

$$\bigcap_{\varphi\in\Phi} \mathrm{Min}_{\leqslant_w}([\![\varphi]\!]) \subseteq \bigcap_{\varphi\in\Phi} [\![B\varphi]\!], \tag{2.12}$$

which with (2.11) and (2.9) implies $\mathrm{Min}_{\leqslant_w}([\![\psi]\!]) \subseteq [\![B\psi]\!]$. Thus, adherence and sensitivity hold for ψ at w, so $\mathcal{M}, w \vDash_n K\psi$ given $\mathcal{M}, w \vDash_n B\psi$.

If (2.7) holds for $x := s$, then by the truth definition (Def. 2.7),

$$\mathcal{M}, w \vDash_s \bigwedge_{\varphi\in\Phi} B\varphi \text{ and } \bigcap_{\varphi\in\Phi} \mathrm{Min}_{\leqslant_w}([\![B\varphi]\!]) \subseteq \bigcap_{\varphi\in\Phi} [\![\varphi]\!]. \tag{2.13}$$

By (2.9), the first conjunct of (2.13) implies $\mathcal{M}, w \vDash_s B\psi$. Given $w \in \mathrm{Min}_{\leqslant_w}(W)$ (Def. 2.2.3b), it follows that $\mathrm{Min}_{\leqslant_w}([\![B\psi]\!]) \subseteq \mathrm{Min}_{\leqslant_w}(W)$ and therefore

$$\mathrm{Min}_{\leqslant_w}([\![B\psi]\!]) \subseteq \bigcap_{\varphi\in\Phi} \mathrm{Min}_{\leqslant_w}([\![B\varphi]\!]) \tag{2.14}$$

by (2.9). Finally, from (d) we have

$$\bigcap_{\varphi\in\Phi}[\![\varphi]\!]\subseteq[\![\psi]\!], \tag{2.15}$$

which with (3.3.2) and the second conjunct of (2.13) implies the safety condition that $\mathrm{Min}_{\leqslant_w}([\![B\psi]\!])\subseteq[\![\psi]\!]$, so $\mathcal{M},w\vDash_s K\psi$ given $\mathcal{M},w\vDash_s B\psi$. □

Remark 2.7 (Dropping Well-Foundedness)**.** We can drop the assumption of well-foundedness used in the above proofs, provided we modify the truth definitions accordingly. For example (cf. Lewis 1973, §2.3), we may define

$$\mathcal{M},w\vDash_{d'} K\varphi \text{ iff } \begin{cases}[\![\varphi]\!]_{d'}=W_w \text{ or}\\ \exists v\in\overline{[\![\varphi]\!]}_{d'}\cap W_w\ \forall u\in\overline{[\![\varphi]\!]}_{d'}\text{: if } u\preceq_w v \text{ then } w\not\rightarrow u,\end{cases} \tag{2.16}$$

which is equivalent to the clause in Definition 2.5 over total well-founded models. I will give the proof for Theorem 2.1.2 that (c) implies the validity of $\chi_{n,m}$ over total RA models according to (2.16). Assume that (2.7) holds for $x:=d'$. If $[\![\varphi]\!]=W_w$ for all $\varphi\in\Phi$, then by (c), $[\![\psi]\!]=W_w$ and hence $\mathcal{M},w\vDash_{d'} K\psi$ for all $\psi\in\Psi$. Otherwise, for every $\varphi\in\Phi$ for which the second case of (2.16) holds, let v_φ be a witness to the existential quantifier. Since $\{v_\varphi\mid\varphi\in\Phi\}$ is finite and nonempty, $\mathrm{Min}_{\preceq_w}(\{v_\varphi\mid\varphi\in\Phi\})$ is nonempty. Consider some $v\in\mathrm{Min}_{\preceq_w}(\{v_\varphi\mid\varphi\in\Phi\})$. Given that $\preceq_w$ is a total preorder,

$$\forall u\in\bigcup_{\varphi\in\Phi}\overline{[\![\varphi]\!]}_{d'}\text{: if } u\preceq_w v \text{ then } w\not\rightarrow u. \tag{2.17}$$

Since $v\in\overline{[\![\varphi]\!]}$ for some $\varphi\in\Phi$, by (c) it follows that $v\in\overline{[\![\psi]\!]}$ for some $\psi\in\Psi$. Now observe that for all $u\in\overline{[\![\psi]\!]}$, $u\preceq_w v$ implies $w\not\rightarrow u$. For if $u\in\overline{[\![\psi]\!]}$, then by (c), $u\in\overline{[\![\varphi]\!]}$ for some $\varphi\in\Phi$, in which case $u\preceq_w v$ implies $w\not\rightarrow u$ by (2.17). Hence v is a witness to the existential in (2.16) for $K\psi$, whence $\mathcal{M},w\vDash_{d'} K\psi$.

We leave the other cases without well-foundedness to the reader.[35]

[35]For H-semantics without well-foundedness (but with totality), define a new $\vDash_{h'}$ relation as in (2.16) but with $\mathcal{M},u\nvDash_{h'} B\varphi$ in place of $w\not\rightarrow u$ and with the belief condition for knowledge. Then the proof of the 'if' direction of Theorem 2.1.4 for $\vDash_{h'}$ is similar to the proof above for $\vDash_{d'}$, but replacing (c) by (d) and replacing $w\not\rightarrow u$ in (2.17) by $\mathcal{M},u\nvDash_{h'} B\psi$, which follows from $\mathcal{M},u\nvDash_{h'} B\varphi$ for

If we do not assume totality, then for D-semantics we modify the right hand side of (2.16) as follows:

$$\forall x(x \in \overline{[\![\varphi]\!]}_{d''} \Rightarrow \exists v \in \overline{[\![\varphi]\!]}_{d''}(v \preceq_w x \;\&\; \forall u \in \overline{[\![\varphi]\!]}_{d''}(u \preceq_w v \Rightarrow w \not\hookrightarrow u))). \tag{2.18}$$

I will now give the proof for Theorem 2.1.3 that (d) implies the validity of $\chi_{n,m}$ over RA models according to this modified D$''$-semantics. For simplicity, let us work out the case where $\vDash_{d''} \varphi_1 \wedge \varphi_2 \leftrightarrow \psi$, which shows the pattern for the general case. Given $\mathcal{M}, w \vDash_{d''} K\varphi_1 \wedge K\varphi_2$, by the truth definition we have:

$$\forall x(x \in \overline{[\![\varphi_1]\!]}_{d''} \Rightarrow \exists v \in \overline{[\![\varphi_1]\!]}_{d''}(v \preceq_w x \;\&\; \forall u \in \overline{[\![\varphi_1]\!]}_{d''}(u \preceq_w v \Rightarrow w \not\hookrightarrow u))); \tag{2.19}$$

$$\forall x(x \in \overline{[\![\varphi_2]\!]}_{d''} \Rightarrow \exists v \in \overline{[\![\varphi_2]\!]}_{d''}(v \preceq_w x \;\&\; \forall u \in \overline{[\![\varphi_2]\!]}_{d''}(u \preceq_w v \Rightarrow w \not\hookrightarrow u))). \tag{2.20}$$

Suppose for contradiction that $\mathcal{M}, w \nvDash_{d''} K\psi$, so by the truth definition,

$$\exists x \in \overline{[\![\psi]\!]}_{d''} \forall v \in \overline{[\![\psi]\!]}_{d''}(v \preceq_w x \Rightarrow \exists u \in \overline{[\![\psi]\!]}_{d''}(u \preceq_w v \;\&\; w \rightarrowtriangle u)). \tag{2.21}$$

Given $x \in \overline{[\![\psi]\!]}_{d''}$ and $\vDash_{d''} \varphi_1 \wedge \varphi_2 \leftrightarrow \psi$, we have $x \in \overline{[\![\varphi_i]\!]}_{d''}$ for some $i \in \{1, 2\}$. Without loss of generality, suppose $i = 1$. Thus, by (2.19),

$$\exists v \in \overline{[\![\varphi_1]\!]}_{d''}(v \preceq_w x \;\&\; \forall u \in \overline{[\![\varphi_1]\!]}_{d''}(u \preceq_w v \Rightarrow w \not\hookrightarrow u)). \tag{2.22}$$

Given $v \in \overline{[\![\varphi_1]\!]}_{d''}$ and $\vDash_{d''} \varphi_1 \wedge \varphi_2 \leftrightarrow \psi$, we have $v \in \overline{[\![\psi]\!]}_{d''}$. It follows by (2.21) that

$$\exists u \in \overline{[\![\psi]\!]}_{d''}(u \preceq_w v \;\&\; w \rightarrowtriangle u). \tag{2.23}$$

Since $u \preceq_w v$ and $w \rightarrowtriangle u$, it follows by (2.22) that $u \notin \overline{[\![\varphi_1]\!]}_{d''}$. Then given $u \in \overline{[\![\psi]\!]}_{d''}$

any $\varphi \in \Phi$ by (d) and Fact 2.7. Finally, since Definition 2.2.3b implies that $\mathrm{Min}_{\leqslant_w}(W) \neq \emptyset$ even if $\leqslant_w$ is not well-founded, it follows from Observation 2.1 that the adherence and safety conditions of N/S-semantics do not require well-foundedness.

and $\vDash_{d''} \varphi_1 \wedge \varphi_2 \leftrightarrow \psi$, we have $u \in \overline{[\![\varphi_2]\!]}_{d''}$. Thus, by (2.20),

$$\exists v' \in \overline{[\![\varphi_2]\!]}_{d''}(v' \preceq_w u \;\&\; \forall u' \in \overline{[\![\varphi_2]\!]}_{d''}(u' \preceq_w v' \Rightarrow w \not\rightarrow u')). \tag{2.24}$$

Given $v' \in \overline{[\![\varphi_2]\!]}_{d''}$ and $\vDash_{d''} \varphi_1 \wedge \varphi_2 \leftrightarrow \psi$, we have $v' \in \overline{[\![\psi]\!]}_{d''}$; and given $v' \preceq_w u \preceq_w v \preceq_w x$ and the transitivity of $\preceq_w$, $v' \preceq_w x$. It follows by (2.21) that

$$\exists u' \in \overline{[\![\psi]\!]}_{d''}(u' \preceq_w v' \;\&\; w \rightarrow u'). \tag{2.25}$$

Since $u' \preceq_w v'$ and $w \rightarrow u'$, it follows by (2.24) that $u' \notin \overline{[\![\varphi_2]\!]}_{d''}$. Given $u' \preceq_w v' \preceq_w u \preceq_w v$ and the transitivity of $\preceq_w$, $u' \preceq_w v$. Then since $u' \preceq_w v$ and $w \rightarrow u'$, it follows by (2.22) that $u' \notin \overline{[\![\varphi_1]\!]}_{d''}$. But together $u' \in \overline{[\![\psi]\!]}_{d''}$, $u' \notin \overline{[\![\varphi_2]\!]}_{d''}$, and $u' \notin \overline{[\![\varphi_2]\!]}_{d''}$ contradict the assumption that $\vDash_{d''} \varphi_1 \wedge \varphi_2 \leftrightarrow \psi$. Hence $\mathcal{M}, w \vDash_{d''} K\psi$.

For H-semantics without totality, we modify the right hand side of (2.16) as follows: $\mathcal{M}, w \vDash_{h''} B\varphi$ and

$$\forall x(x \in \overline{[\![\varphi]\!]}_{h''} \Rightarrow \exists v \in \overline{[\![\varphi]\!]}_{h''}(v \preceq_w x \;\&\; \forall u \in \overline{[\![\varphi_1]\!]}_{h''}(u \preceq_w v \Rightarrow \mathcal{M}, u \nvDash_{h''} B\varphi))). \tag{2.26}$$

The proof for Theorem 2.1.4 that (d) implies the validity of $\chi_{n,m}$ over CB models according to this modified H″-semantics follows the same pattern as the proof for D″-semantics above, only with the additional use of Fact 2.7.

2.6.2 Completeness for Total RA Models

We turn now to the 'only if' directions of Theorem 2.1. The proof for part 1 of the theorem, which we omit, is a much simpler application of the general approach used for the other parts. In this section, we treat the 'only if' direction of part 2. This is the most involved part of the proof and takes us most of the way toward the 'only if' direction of part 3, treated in §2.6.3. It may help at times to recall the proof sketch given after Fact 2.6 above.

In §2.6.2, I define what it is for the $\chi_{n,m}$ in Theorem 2.1 to be *T-unpacked*. In §2.6.2, I show that if a T-unpacked $\chi_{n,m}$ does not satisfy (a) or (c) of Theorem 2.1,

then it is falsified by a finite *total* RA model according to D-semantics. In fact, it is falsified by a finite *linear* RA model with the *universal field* property (Def. 2.3.4). Finally, in §2.6.2 we give upper bounds on the size of and complexity of finding falsifying models in Corollaries 2.2 and 2.3.

T-unpacking Formulas

Toward defining what it is for $\chi_{n,m}$ (Notation 2.3) to be T-unpacked, let us first define a normal form for the $\varphi_1, \ldots, \varphi_n$ in $\chi_{n,m}$. For our purposes, we need only define the normal form for the top (propositional) level of each φ_i.

Definition 2.8 (DNF)**.** A formula in the epistemic language is in (propositional) *disjunctive normal form* (DNF) iff it is of the form

$$\bigvee(\alpha \wedge \bigwedge K\beta \wedge \bigwedge \neg K\gamma),$$

where α is propositional (a conjunction of literals, but it will not matter here), and β and γ are any formulas.

Roughly speaking, we T-unpack a conditional $\chi_{n,m}$ by using the T axiom, $K\varphi_i \rightarrow \varphi_i$, to replace $K\varphi_i$ in the antecedent with the equivalent $\varphi_i \wedge K\varphi_i$ and then use propositional logic to put φ_i in its appropriate place; e.g., if φ_i is $\neg K\gamma$, then we move $K\gamma$ to the consequent to become one of the $K\psi$'s. After the following general definition and result, we work out a concrete example.

Definition 2.9 (T-unpacked)**.** For any (possibly empty) sequence of formulas $\psi_1, \ldots, \psi_m$, a formula of the form $\chi_{0,m}$ is T-unpacked; and for φ_{n+1} in DNF, a formula of the form $\chi_{n+1,m}$ is T-unpacked iff $\chi_{n,m}$ is T-unpacked and there is a disjunct δ of φ_{n+1} such that:

1. the α conjunct in δ is a conjunct of φ_0;
2. for all $K\beta$ conjuncts in δ, there is some $i \leq n$ such that $\varphi_i = \beta$;
3. for all $\neg K\gamma$ conjuncts in δ, there is some $j \leq m$ such that $\psi_j = \gamma$.

The following proposition will be used to prove several later results.

Proposition 2.2 (T-unpacking)**.** Every formula in the epistemic language is equivalent over RA models in C/D/L-semantics (and over CB models in H/N/S-semantics) to a conjunction of T-unpacked formulas of the form $\chi_{n,m}$.

Proof. By propositional logic, every formula θ is equivalent to a conjunction of formulas of the conditional (disjunctive) form $\chi_{n,m}$. Also by propositional logic, every φ_i in the antecedent of $\chi_{n,m}$ can be converted into an equivalent $\varphi_i^\vee$ in DNF; and since φ_i and $\varphi_i^\vee$ are equivalent, so are $K\varphi_i$ and $K\varphi_i^\vee$ by the semantics. To obtain an equivalent of θ in which each $\chi_{n,m}$ is T-unpacked, we repeatedly use the following equivalences, easily derived using propositional logic and the valid T axiom, $K\psi \to \psi$. Where ζ and η are any formulas,

$$\begin{aligned}
& \zeta \wedge K\big(\bigvee_{k\le l}\delta_k\big) \to \eta \\
\Leftrightarrow\ & \zeta \wedge \big(\bigvee_{k\le l}\delta_k\big) \wedge K\big(\bigvee_{k\le l}\delta_k\big) \to \eta \\
\Leftrightarrow\ & \bigwedge_{k\le l}\Big(\zeta \wedge \delta_k \wedge K\big(\bigvee_{k\le l}\delta_k\big) \to \eta\Big) \\
\Leftrightarrow\ & \bigwedge_{k\le l}\Big(\zeta \wedge \alpha^k \wedge \bigwedge K\beta^k \wedge K\big(\bigvee_{k\le l}\delta_k\big) \to \eta \vee \bigvee K\gamma^k\Big),
\end{aligned}$$

where each δ_k is of the form $\alpha^k \wedge \bigwedge K\beta^k \wedge \bigwedge \neg K\gamma^k$. Compare conditions 1 - 3 of Definition 2.9 to the relation of δ_k to the k-th conjunct in the last line. □

Example 2.2 (T-unpacking cont.)**.** Let us T-unpack the following formula:

$$\underbrace{K\Big(\underbrace{\big(K\underbrace{(Kp \vee q)}_{\beta_1^1} \wedge K\underbrace{\neg Kq}_{\beta_2^1} \wedge \neg K\underbrace{Kr}_{\gamma_1^1}\big)}_{\delta_1} \vee \underbrace{K\underbrace{\neg Kr}_{\beta_1^2}}_{\delta_2}\Big)}_{\varphi} \to K\psi.$$

No matter what we substitute for ψ, the form of the final result will be the same, since T-unpacking does nothing to formulas already in the consequent.

As in the proof of Proposition 2.2, we derive a string of equivalences, obtaining formulas in boldface by applications of the T axiom and otherwise using only propositional logic:

$K\varphi \to K\psi \Leftrightarrow \boldsymbol{\varphi} \wedge K\varphi \to K\psi$;

then since φ is a disjunction, we split into two conjuncts:

$$\Leftrightarrow (\delta_1 \wedge K\varphi \rightarrow K\psi) \wedge (\delta_2 \wedge K\varphi \rightarrow K\psi);$$

then we move the negated $K\gamma_1^1$ in δ_1 to the first consequent and rewrite as

$$\Leftrightarrow (K\beta_1^1 \wedge K\beta_2^1 \wedge K\varphi \rightarrow K\psi \vee K\gamma_1^1) \wedge (K\beta_1^2 \wedge K\varphi \rightarrow K\psi);$$

then we apply the T axiom to the $K\beta$ formulas:

$$\Leftrightarrow (\boldsymbol{\beta_1^1} \wedge \boldsymbol{\beta_2^1} \wedge K\beta_1^1 \wedge K\beta_2^1 \wedge K\varphi \rightarrow K\psi \vee K\gamma_1^1) \wedge (\boldsymbol{\beta_1^2} \wedge K\beta_1^2 \wedge K\varphi \rightarrow K\psi);$$

then we move the negated Kq in β_2^1 and Kr in β_1^2 to the consequents:

$$\Leftrightarrow (\beta_1^1 \wedge K\beta_1^1 \wedge K\beta_2^1 \wedge K\varphi \rightarrow K\psi \vee K\gamma_1^1 \vee Kq) \wedge (K\beta_1^2 \wedge K\varphi \rightarrow K\psi \vee Kr);$$

since β_1^1 is another disjunction, we split the first conjunct into two:

$$\begin{aligned}\Leftrightarrow\ & (Kp \wedge K\beta_1^1 \wedge K\beta_2^1 \wedge K\varphi \rightarrow K\psi \vee K\gamma_1^1 \vee Kq) \wedge \\ & (q \wedge K\beta_1^1 \wedge K\beta_2^1 \wedge K\varphi \rightarrow K\psi \vee K\gamma_1^1 \vee Kq) \wedge \\ & (K\beta_1^2 \wedge K\varphi \rightarrow K\psi \vee Kr);\end{aligned}$$

finally, we apply the T axiom to Kp and rewrite as:

$$\begin{aligned}\Leftrightarrow\ & (\underline{\boldsymbol{p}}_{\varphi_0} \wedge K\underline{p}_{\varphi_1} \wedge K\underline{(Kp \vee q)}_{\varphi_2} \wedge K\underline{\neg Kq}_{\varphi_3} \wedge K\underline{\varphi}_{\varphi_4} \rightarrow K\underline{\psi}_{\psi_1} \vee K\underline{Kr}_{\psi_2} \vee K\underline{q}_{\psi_3}) \\ & \wedge (\underline{q}_{\varphi_0'} \wedge K\underline{(Kp \vee q)}_{\varphi_1'} \wedge K\underline{\neg Kq}_{\varphi_2'} \wedge K\underline{\varphi}_{\varphi_3'} \rightarrow K\underline{\psi}_{\psi_1'} \vee K\underline{Kr}_{\psi_2'} \vee K\underline{q}_{\psi_3'}) \\ & \wedge (K\underline{\neg Kr}_{\varphi_1''} \wedge K\underline{\varphi}_{\varphi_2''} \rightarrow K\underline{\psi}_{\psi_1''} \vee K\underline{r}_{\psi_2''}).\end{aligned}$$

Observe that the three conjuncts are T-unpacked according to Definition 2.9.

Countermodel Construction

Our approach to proving the 'only if' direction of Theorem 2.1.2 is to assume that (a) and (c) fail, from which we infer the existence of models that can be "glued together" to construct a countermodel for $\chi_{n,m}$. For a clear illustration of this approach applied to basic modal models with arbitrary accessibility relations, see van Benthem 2010, §4.3. There are two important differences in what we must do here. First, since we are dealing with reflexive models in which $K\varphi \rightarrow \varphi$ is valid, we must use T-unpacking.

Second, since we are dealing with a hybrid of relational and *ordering* semantics, we cannot simply glue all of the relevant models together at once, as in the basic modal case; instead, we must put them in the right order, which we do inductively.

The construction has two main parts. First, we inductively build up a kind of "pre-model" that falsifies $\chi_{n,m}$. Second, assuming that $\chi_{n,m}$ is T-unpacked, we can then convert the pre-model into an RA model that falsifies $\chi_{n,m}$.

Definition 2.10 (Pre-Model)**.** A pointed *pre-model* is a pair $\mathcal{M}, v$, with $\mathcal{M} = \langle W, \rightarrow\!\!\triangleright, \preceq, V\rangle$ and $v \in W$, where W, $\rightarrow\!\!\triangleright$, $\preceq_w$ for $w \in W \setminus \{v\}$, and V are as in Definition 2.2; $\preceq_v$ satisfies Definition 2.2.3a, but for all $w \in W$, $v \notin W_w$.

Hence a pointed pre-model is not a pointed RA model, since Definition 2.2.3b requires that $v \in W_v$ for an RA model. However, truth at a pointed pre-model is defined in the same way as truth at a pointed RA model in Definition 2.5.

The following lemma shows how we will build up our model in the inductive construction of Lemma 2.3. It is important to note that Lemmas 2.2 and 2.3 hold for any $\chi_{n,m}$ as in Notation 2.3, whether or not it is T-unpacked.

Lemma 2.2 (Pre-Model Extension)**.** Assume there is a linear pointed pre-model $\mathcal{M}, w$ such that $\mathcal{M}, w \not\models_d \chi_{n,m}$.

1. If $\psi_1 \wedge \cdots \wedge \psi_m \rightarrow \varphi_{n+1}$ is not D-valid over linear RA models, then there is a linear pointed pre-model $\mathcal{M}^\sharp, w$ such that $\mathcal{M}^\sharp, w \not\models_d \chi_{n+1,m}$.

2. If $\varphi_1 \wedge \cdots \wedge \varphi_n \rightarrow \psi_{m+1}$ is not D-valid over linear RA models, then there is a linear pointed pre-model $\mathcal{M}^\flat, w$ such that $\mathcal{M}^\flat, w \not\models_d \chi_{n,m+1}$.

Proof. For part 1, let $\mathcal{N} = \langle N, \rightarrow\!\!\triangleright^{\mathcal{N}}, \preceq^{\mathcal{N}}, V^{\mathcal{N}}\rangle$ with $v \in N$ be a linear RA model such that $\mathcal{N}, v \not\models_d \psi_1 \wedge \cdots \wedge \psi_m \rightarrow \varphi_{n+1}$. By assumption, there is a linear pre-model $\mathcal{M} = \langle M, \rightarrow\!\!\triangleright^{\mathcal{M}}, \preceq^{\mathcal{M}}, V^{\mathcal{M}}\rangle$ with point $w \in M$ such that $\mathcal{M}, w \not\models_d \chi_{n,m}$. Define $\mathcal{M}^\sharp = \langle W^\sharp, \rightarrow\!\!\triangleright^\sharp, \preceq^\sharp, V^\sharp\rangle$ as follows (see Fig. 2.4):

$W^\sharp = M \cup N$ (we can assume $M \cap N = \emptyset$); $\rightarrow\!\!\triangleright^\sharp = \rightarrow\!\!\triangleright^{\mathcal{M}} \cup \rightarrow\!\!\triangleright^{\mathcal{N}}$;

$\preceq^\sharp_w = \preceq^{\mathcal{M}}_w \cup \{\langle v, x\rangle \mid x = v \text{ or } x \in M_w\}$, where M_w is the field of $\preceq^{\mathcal{M}}_w$;

$\preceq_x^\sharp = \preceq_x^{\mathcal{M}}$ for all $x \in M \setminus \{w\}$; $\preceq_y^\sharp = \preceq_y^{\mathcal{N}}$ for all $y \in N$;

$V^\sharp(p) = V^{\mathcal{M}}(p) \cup V^{\mathcal{N}}(p)$.

Observe that $\mathcal{M}^\sharp, w$ is a linear pointed pre-model.

It is easy to verify that for all formulas ξ and $x \in M \setminus \{w\}$,

$$\mathcal{M}^\sharp, x \vDash_d \xi \text{ iff } \mathcal{M}, x \vDash_d \xi \text{ ; and } \mathcal{M}^\sharp, v \vDash_d \xi \text{ iff } \mathcal{N}, v \vDash_d \xi. \tag{2.27}$$

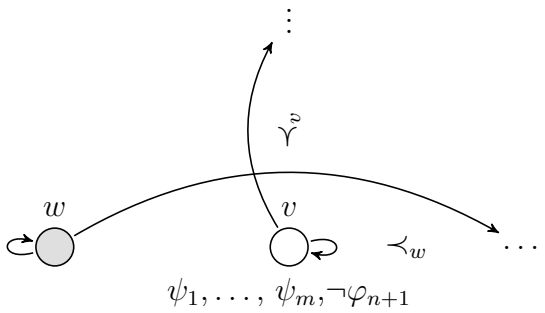

Figure 2.4: part of the extended pre-model $\mathcal{M}^\sharp$ for Lemma 2.2.1

Given $\mathcal{M}, w \nvDash_d \chi_{n,m}$ and the truth definition (Def. 2.5),

$$\bigcup_{1 \le i \le n} \mathrm{Min}_{\preceq_w^{\mathcal{M}}}(\overline{[\![\varphi_i]\!]}^{\mathcal{M}}) \cap \mathrel{-\!\!\triangleright}^{\mathcal{M}}(w) = \emptyset. \tag{2.28}$$

It follows by the construction of $\mathcal{M}^\sharp$ and (2.27) that

$$\bigcup_{1 \le i \le n+1} \mathrm{Min}_{\preceq_w^\sharp}(\overline{[\![\varphi_i]\!]}^{\mathcal{M}^\sharp}) \cap \mathrel{-\!\!\triangleright}^\sharp(w) = \emptyset, \tag{2.29}$$

which is equivalent to $\mathcal{M}^\sharp, w \vDash_d K\varphi_1 \wedge \cdots \wedge K\varphi_{n+1}$ by the truth definition. The construction of $\mathcal{M}^\sharp$ and (2.27) also guarantee that for all $k \le m$,

$$\mathrm{Min}_{\preceq_w^{\mathcal{M}}}(\overline{[\![\psi_k]\!]}^{\mathcal{M}}) \cap \mathrel{-\!\!\triangleright}^{\mathcal{M}}(w) \subseteq \mathrm{Min}_{\preceq_w^\sharp}(\overline{[\![\psi_k]\!]}^{\mathcal{M}^\sharp}) \cap \mathrel{-\!\!\triangleright}^\sharp(w). \tag{2.30}$$

Given $\mathcal{M}, w \nvDash_d \chi_{n,m}$, for all $k \le m$ the left side of (2.30) is nonempty, so the right

side is nonempty. Hence by the truth definition, $\mathcal{M}^\sharp, w \not\models_d K\psi_1 \vee \cdots \vee K\varphi_m$. Finally, since φ_0 is propositional, $\mathcal{M}, w \models \varphi_0$ implies $\mathcal{M}^\sharp, w \models \varphi_0$ by definition of $V^\sharp$. It follows from the preceding facts that $\mathcal{M}^\sharp, w \not\models_d \chi_{n+1,m}$.

For part 2, let $\mathcal{O} = \langle O, \rightarrow\!\!\triangleright^{\mathcal{O}}, \preceq^{\mathcal{O}}, V^{\mathcal{O}}\rangle$ with $u \in O$ be a linear RA model such that $\mathcal{O}, u \not\models_d \varphi_1 \wedge \cdots \wedge \varphi_n \rightarrow \psi_{m+1}$. Given $\mathcal{M}, w$ as in part 1, define $\mathcal{M}^\flat = \langle W^\flat, \rightarrow\!\!\triangleright^\flat, \preceq^\flat, V^\flat\rangle$ from $\mathcal{M}$ and $\mathcal{O}$ in the same way as we defined $\mathcal{M}^\sharp$ from $\mathcal{M}$ and $\mathcal{N}$ for part 1, except that $\rightarrow\!\!\triangleright^\flat = \rightarrow\!\!\triangleright^{\mathcal{M}} \cup \rightarrow\!\!\triangleright^{\mathcal{O}} \cup \{w, u\}$ (see Fig. 2.5). Observe that $\mathcal{M}^\flat, w$ is a linear pointed pre-model.

It is easy to verify that for all formulas ξ and $x \in M \setminus \{w\}$,

$$\mathcal{M}^\flat, x \models_d \xi \text{ iff } \mathcal{M}, x \models_d \xi\,; \text{ and } \mathcal{M}^\flat, u \models_d \xi \text{ iff } \mathcal{O}, u \models_d \xi. \tag{2.31}$$

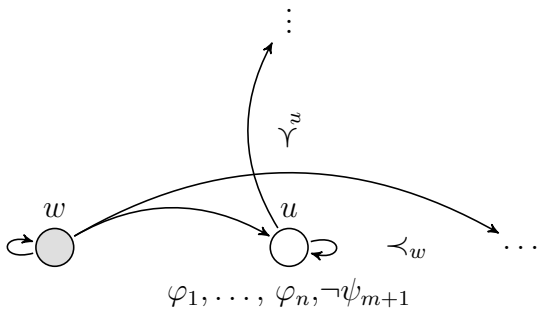

Figure 2.5: part of the extended pre-model $\mathcal{M}^\flat$ for Lemma 2.2.2

As in the proof of part 1, (2.28) holds for $\mathcal{M}$. It follows by the construction of $\mathcal{M}^\flat$ and (2.31) that (2.28) also holds for $\mathcal{M}^\flat$ and $\rightarrow\!\!\triangleright^\flat$ in place of $\mathcal{M}$ and $\rightarrow\!\!\triangleright^{\mathcal{M}}$, so $\mathcal{M}^\flat, w \models_d K\varphi_1 \wedge \cdots \wedge K\varphi_n$ by the truth definition. Also as in the proof of part 1, $\mathrm{Min}_{\preceq^{\mathcal{M}}_w}(\overline{[\![\psi_k]\!]}^{\mathcal{M}}) \cap \rightarrow\!\!\triangleright^{\mathcal{M}}(w)$ is nonempty for all $k \leq m$. It follows by the construction of $\mathcal{M}^\flat$ and (2.31) that $\mathrm{Min}_{\preceq^\flat_w}(\overline{[\![\psi_k]\!]}^{\mathcal{M}^\flat}) \cap \rightarrow\!\!\triangleright^\flat(w)$ is nonempty for all $k \leq m+1$, so $\mathcal{M}^\flat, w \not\models K\psi_1 \vee \cdots \vee K\psi_{m+1}$ by the truth definition. Finally, since φ_0 is propositional, $\mathcal{M}, w \models \varphi_0$ implies $\mathcal{M}^\flat, w \models \varphi_0$ by definition of $V^\flat$. It follows from the preceding facts that $\mathcal{M}^\flat, u \not\models_d \chi_{n,m+1}$. □

Remark 2.8 (Properties of $\rightarrow\!\!\triangleright$)**.** Lemma 2.2 also holds for the class of RA models/

pre-models in which $\rightarrowtriangle$ is an equivalence relation, so that Theorem 2.1.2-3 will as well. For part 1, if $\mathcal{M}$ and $\mathcal{N}$ are in this class, so is $\mathcal{M}^\sharp$, since the union of two disjoint equivalence relations is an equivalence relation. For part 2, suppose $\mathcal{M}$ and $\mathcal{O}$ are in the class. Since we have added an arrow from w to u, $\mathcal{M}^\flat$ may not be in the class. In this case, let $\rightarrowtriangle^+$ be the minimal extension of $\rightarrowtriangle^\flat$ that is an equivalence relation. One can check that by construction of $\mathcal{M}^\flat$, for all $w \in W^\flat$, $(\rightarrowtriangle^+(w) \setminus \rightarrowtriangle^\flat(w)) \cap W_w = \emptyset$. It follows that $\mathcal{M}^\flat$ and $\mathcal{M}^+ = \langle W^\flat, \rightarrowtriangle^+, \preceq^\flat, V^\flat \rangle$ satisfy the same formulas according to D-semantics.

Using Lemma 2.2, we can now carry out our inductive construction.

Lemma 2.3 (Pre-Model Construction)**.** If neither (a) nor (c) of Theorem 2.1 holds for $\chi_{n,m}$, then there is a linear pointed pre-model $\mathcal{M}, w$ such that $\mathcal{M}, w \not\models_d \chi_{n,m}$.

Proof. The proof is by induction on m with a subsidiary induction on n.

Base case for m. Assume that neither (a) nor (c) holds for $\chi_{n,0}$.[36] Let $\mathcal{M} = \langle W, \rightarrowtriangle, \preceq, V \rangle$ be such that $W = \{w\}$, $\rightarrowtriangle = \{\langle w, w \rangle\}$, $\preceq_w = \emptyset$, and V is any valuation such that $\mathcal{M}, w \models \varphi_0$, which exists since (a) does not hold for $\chi_{n,0}$. Then $\mathcal{M}, w$ is a linear pointed pre-model such that $\mathcal{M}, w \not\models_d \chi_{n,0}$.

Inductive step for m. Assume for induction on m that for any $\beta_1, \ldots, \beta_m$ and any n, if neither (a) nor (c) holds for $\chi := \varphi_0 \wedge K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\beta_1 \vee \cdots \vee K\beta_m$, then there is a linear pointed pre-model $\mathcal{M}, w$ with $\mathcal{M}, w \not\models_d \chi$. Assume that for some $\psi_1, \ldots, \psi_{m+1}$, neither (a) nor (c) holds for $\chi_{n,m+1}$. We prove by induction on n that there a linear $\mathcal{M}', w$ with $\mathcal{M}', w \not\models_d \chi_{n,m+1}$.

Base case for n. Assume neither (a) nor (c) holds for $\chi_{0,m+1}$. Since (c) does not hold, for all $j \leq m+1$, $\not\models_d \top \leftrightarrow \psi_j$ and hence $\not\models_d \top \rightarrow \psi_j$. Starting with $\mathcal{M}, w$ defined as in the base case for m such that $\mathcal{M}, w \not\models \chi_{0,0}$, apply Lemma 2.2.2 $m+1$ times to obtain an $\mathcal{M}', w$ with $\mathcal{M}', w \not\models \chi_{0,m+1}$.

Inductive step for n. Assume for induction on n that for any $\alpha_0, \ldots, \alpha_n$, if neither (a) nor (c) holds for $\chi := \alpha_0 \wedge K\alpha_1 \wedge \cdots \wedge K\alpha_n \rightarrow K\psi_1 \vee \cdots \vee K\psi_{m+1}$, then there is a linear pointed pre-model $\mathcal{M}, w$ with $\mathcal{M}, w \not\models_d \chi$. Assume that for some $\varphi_0, \ldots, \varphi_{n+1}$, neither (a) nor (c) holds for $\chi_{n+1,m+1}$.

[36] Recall that $\chi_{n,0}$ is of the form $\varphi_0 \wedge K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow \bot$.

Case 1: $\vDash_d \varphi_1 \wedge \cdots \wedge \varphi_{n+1} \rightarrow \psi_1 \wedge \cdots \wedge \psi_{m+1}$. Then since (c) does not hold for $\chi_{n+1,m+1}$, $\nvDash_d \psi_1 \wedge \cdots \wedge \psi_{m+1} \rightarrow \varphi_1 \wedge \cdots \wedge \varphi_{n+1}$, in which case $\nvDash_d \psi_1 \wedge \cdots \wedge \psi_{m+1} \rightarrow \varphi_i$ for some $i \leq n+1$. Without loss of generality, assume

$$\nvDash_d \psi_1 \wedge \cdots \wedge \psi_{m+1} \rightarrow \varphi_{n+1}. \tag{2.32}$$

Since neither (a) nor (c) holds for $\chi_{n+1,m+1}$, neither holds for $\chi_{n,m+1}$. Hence by the inductive hypothesis for n there is a linear pointed pre-model $\mathcal{M}, w$ such that $\mathcal{M}, w \nvDash_d \chi_{n,m+1}$, which with (2.32) and Lemma 2.2.1 implies that there is a linear pointed pre-model $\mathcal{M}^\sharp, w$ such that $\mathcal{M}^\sharp, w \nvDash_d \chi_{n+1,m+1}$.

Case 2: $\nvDash_d \varphi_1 \wedge \cdots \wedge \varphi_{n+1} \rightarrow \psi_1 \wedge \cdots \wedge \psi_{m+1}$. Then for some $j \leq m+1$, $\nvDash \varphi_1 \wedge \cdots \wedge \varphi_{n+1} \rightarrow \psi_j$. Without loss of generality, assume

$$\nvDash_d \varphi_1 \wedge \cdots \wedge \varphi_{n+1} \rightarrow \psi_{m+1}. \tag{2.33}$$

Since neither (a) nor (c) holds for $\chi_{n+1,m+1}$, neither holds for $\chi_{n+1,m}$. Hence by the inductive hypothesis for m there is a linear pointed pre-model $\mathcal{M}, w$ such that $\mathcal{M}, w \nvDash_d \chi_{n+1,m}$, which with (2.33) and Lemma 2.2.2 implies that there is a linear pointed pre-model $\mathcal{M}^\flat, w$ such that $\mathcal{M}^\flat, w \nvDash_d \chi_{n+1,m+1}$. □

Finally, if $\chi_{n,m}$ is T-unpacked (Def. 2.9), then we can convert the falsifying pre-model obtained from Lemma 2.3 into a falsifying RA model.

Lemma 2.4 (Pre-Model to Model Conversion)**.** Given a linear pointed pre-model $\mathcal{M}, w$ and a T-unpacked $\chi_{n,m}$ such that $\mathcal{M}, w \nvDash_d \chi_{n,m}$, there is a linear pointed RA model $\mathcal{M}^c, w$ such that $\mathcal{M}^c, w \nvDash_d \chi_{n,m}$.

Proof. Where $\mathcal{M} = \langle W, \twoheadrightarrow, \preceq, V\rangle$, define $\mathcal{M}^c = \langle W, \twoheadrightarrow, \preceq^c, V\rangle$ such that for all $v \in W \setminus \{w\}$, $\preceq^c_v = \preceq_v$, and $\preceq^c_w = \preceq_w \cup \{\langle w, v\rangle \mid v \in \{w\} \cup W_w\}$, where W_w is the field of $\preceq_w$. Since w is strictly minimal in $\preceq^c_w$, $\mathcal{M}^c$ is a linear RA model. (Note, however, that w is still not in the field of $\preceq^c_v$ for any $v \in W \setminus \{w\}$.) By construction of $\mathcal{M}^c$, together $\mathcal{M}, w \nvDash_d K\psi_1 \vee \cdots \vee K\psi_m$ and $w \twoheadrightarrow w$ imply

$$\mathcal{M}^c, w \nvDash_d K\psi_1 \vee \cdots \vee K\psi_m. \tag{2.34}$$

We prove by induction that for all $k \leq n$,

$$\mathcal{M}^c, w \vDash_d \varphi_0 \wedge K\varphi_1 \wedge \cdots \wedge K\varphi_k. \tag{2.35}$$

The base case of $k = 0$ is immediate since φ_0 is propositional, $\mathcal{M}, w \vDash \varphi_0$, and $\mathcal{M}$ and $\mathcal{M}^c$ have the same valuations. Assuming (2.35) holds for $k < n$, we must show $\mathcal{M}^c, w \vDash_d K\varphi_{k+1}$. Since $\chi_{n,m}$ is T-unpacked, together Definition 2.9, (2.34), and (2.35) imply $\mathcal{M}^c, w \vDash_d \varphi_{k+1}$. Since $\mathcal{M}, w \vDash_d K\varphi_{k+1}$, we have $\mathrm{Min}_{\preceq_w}(\overline{[\![\varphi_{k+1}]\!]}^{\mathcal{M}}) \cap \twoheadrightarrow(w) = \emptyset$ by the truth definition (Def. 2.5). It follows, given the construction of $\mathcal{M}^c$ and the fact that $\mathcal{M}^c, w \vDash_d \varphi_{k+1}$, that $\mathrm{Min}_{\preceq^c_w}(\overline{[\![\varphi_{k+1}]\!]}^{\mathcal{M}^c}) \cap \twoheadrightarrow(w) = \emptyset$, which gives $\mathcal{M}^c, w \vDash_d K\varphi_{k+1}$, as desired. □

The proof of the 'only if' direction of Theorem 2.1.2 is complete. By Lemmas 2.3 and 2.4, if a T-unpacked $\chi_{n,m}$ does not satisfy (a) or (c) of Theorem 2.1, then it is falsified by a linear—and hence total—RA model according to D-semantics. Indeed, as the next proposition and Corollary 2.2 together show, it is falsified by an RA model with the *universal field* property (Def. 2.3.4).

Proposition 2.3 (Universalization)**.** Where $\mathcal{M} = \langle W, \twoheadrightarrow, \preceq, V\rangle$ is a finite RA model, there is a finite RA model $\mathcal{M}^u = \langle W^u, \twoheadrightarrow^u, \preceq^u, V^u\rangle$ with the universal field property, such that $W \subseteq W^u$ and for all $w \in W$ and all φ,

$$\mathcal{M}, w \vDash_d \varphi \text{ iff } \mathcal{M}^u, w \vDash_d \varphi.$$

If $\mathcal{M}$ is total, $\mathcal{M}^u$ is also total. If $\mathcal{M}$ is linear, $\mathcal{M}^u$ is also linear.

Proof. Given $\mathcal{M} = \langle W, \twoheadrightarrow, \preceq, V\rangle$, suppose that for some $w, v \in W$, $v \notin W_w$, so $v \neq w$. Define $\mathcal{M}' = \langle W', \twoheadrightarrow', \preceq', V'\rangle$ such that $W' = W$; $\twoheadrightarrow' = \twoheadrightarrow \setminus \{\langle w, v\rangle\}$; $\preceq'_w = \preceq_w \cup \{\langle x, v\rangle \mid x \in W_w \cup \{v\}\}$; $\preceq'_y = \preceq_y$ for $y \in W \setminus \{w\}$; and $V' = V$. In other words, v becomes the least relevant world at w and eliminated at w in $\mathcal{M}'$. Given $v \notin W_w$, one can show by induction on φ that for all $x \in W$, $\mathcal{M}, x \vDash_d \varphi$ iff $\mathcal{M}', x \vDash_d \varphi$. Applying the transformation $\mathcal{M} \mapsto \mathcal{M}'$ successively no more than $|W|^2$ times with other pairs of worlds like w and v yields a model $\mathcal{M}^u$ with the universal field property. If $\mathcal{M}$ is total/linear, so is $\mathcal{M}^u$.

If we require that $\rightarrow\!\!\triangleright$ be an equivalence relation, then the transformation above will not work in general, since we may lose transitivity or symmetry by setting $w \not\rightarrow\!\!\triangleright' v$. To solve this problem, we first make an isomorphic copy of $\mathcal{M}$, labeled $\mathcal{M}^\star = \langle W^\star, \rightarrow\!\!\triangleright^\star, \preceq^\star, V^\star\rangle$. For every $w \in W$, let $w^\star$ be its isomorphic copy in $W^\star$. Define $\mathcal{N} = \langle W^{\mathcal{N}}, \rightarrow\!\!\triangleright^{\mathcal{N}}, \preceq^{\mathcal{N}}, V^{\mathcal{N}}\rangle$ as follows: $W^{\mathcal{N}} = W \cup W^\star$; $\rightarrow\!\!\triangleright^{\mathcal{N}} = \rightarrow\!\!\triangleright \cup \rightarrow\!\!\triangleright^\star$; $V^{\mathcal{N}}(p) = V(p) \cup V^\star(p)$; for all $w \in W$, $\preceq^{\mathcal{N}}_w = \preceq_w \cup \{\langle v, u\rangle \mid v \in W^{\mathcal{N}} \text{ and } u \in W^\star\}$; for all $w^\star \in W^\star$, $\preceq^{\mathcal{N}}_{w^\star} = \preceq_{w^\star} \cup \{\langle v, u\rangle \mid v \in W^{\mathcal{N}} \text{ and } u \in W\}$. In other words, $\mathcal{N}$ is the result of first taking the disjoint union of $\mathcal{M}$ and $\mathcal{M}^\star$ (so there are no $v \in W$ and $u \in W^\star$ such that $v \rightarrow\!\!\triangleright^{\mathcal{N}} u$ or $u \rightarrow\!\!\triangleright^{\mathcal{N}} v$) and then making all worlds in $W^\star$ the least relevant worlds from the perspective of all worlds in W, and vice versa.[37] Given this construction, it is easy to prove by induction that for all $w \in W$ and formulas φ, $\mathcal{M}, w \vDash_d \varphi$ iff $\mathcal{N}, w \vDash_d \varphi$ iff $\mathcal{N}, w^\star \vDash_d \varphi$. Moreover, $\rightarrow\!\!\triangleright^{\mathcal{N}}$ is an equivalence relation if $\rightarrow\!\!\triangleright$ is.

Next we turn $\mathcal{N}$ into a model with universal fields, without changing $\rightarrow\!\!\triangleright^{\mathcal{N}}$. Suppose that for $w, v \in W$, v is not in the field of $\preceq^{\mathcal{N}}_w$, which is the case iff $v^\star$ is not in the field of $\preceq^{\mathcal{N}}_{w^\star}$. (Remember that for all $w \in W$ and $u \in W^\star$, u is in the field of $\preceq^{\mathcal{N}}_w$ and vice versa.) Let $\mathcal{N}' = \langle W', \rightarrow\!\!\triangleright', \preceq', V'\rangle$ be such that: $W' = W^{\mathcal{N}}$; $\rightarrow\!\!\triangleright' = \rightarrow\!\!\triangleright^{\mathcal{N}}$; $V' = V^{\mathcal{N}}$; for all $u \in W' \setminus \{w, w^\star\}$, $\preceq'_u = \preceq^{\mathcal{N}}_u$; $\preceq'_w = \preceq^{\mathcal{N}}_w \cup \{\langle x, v\rangle \mid x \in W^{\mathcal{N}}_w \cup \{v\}\}$; and $\preceq'_{w^\star} = \preceq^{\mathcal{N}}_{w^\star} \cup \{\langle x, v^\star\rangle \mid x \in W^{\mathcal{N}}_{w^\star} \cup \{v^\star\}\}$. It follows that for all $x \in W^{\mathcal{N}}_w$, $x \preceq'_w v^\star \prec'_w v$; and for all $x \in W^{\mathcal{N}}_{w^\star}$, $x \preceq'_{w^\star} v \prec'_{w^\star} v^\star$. Since $w \not\rightarrow\!\!\triangleright' v^\star$ and $w^\star \not\rightarrow\!\!\triangleright' v$, one can prove by induction that for all φ and $u \in W$, $\mathcal{N}, u \vDash_d \varphi$ iff $\mathcal{N}', u \vDash_d \varphi$ iff $\mathcal{N}', u^\star \vDash_d \varphi$. The key is that although we put v in the field of $\preceq'_w$, this cannot make any $K\psi$ formula that is true at $\mathcal{N}, w$ false at $\mathcal{N}', w$, for if $\mathcal{N}', v \nvDash_d \psi$, then by the inductive hypothesis $\mathcal{N}', v^\star \nvDash_d \psi$, and $v^\star$ is more relevant than v and eliminated at w; similarly, although we put $v^\star$ in the field of $\preceq'_{w^\star}$, this cannot make any $K\psi$ formula that is true at $\mathcal{N}, w^\star$ false at $\mathcal{N}', w^\star$. Applying the transformation $\mathcal{N} \mapsto \mathcal{N}'$ successively no more than $|W^{\mathcal{N}}|^2$ times with other worlds like w and v yields a universalized $\mathcal{M}^u$. □

[37] If we want to stay within the class of *linear* models, then we must change the definition of $\preceq^{\mathcal{N}}_w$ so that it extends the linear order $\preceq_w$ with an arbitrary linear order on $W^\star$ that makes all worlds in $W^\star$ less relevant than all worlds in W, and similarly for $\preceq^{\mathcal{N}}_{w^\star}$.

Finite Models and Complexity

From the proofs of §2.6.2, we obtain results on finite models and the complexity of satisfiability for D-semantics over total (linear, universal) RA models.

Corollary 2.2 (Effective Finite Model Property)**.** For any formula φ of the epistemic language, if φ is satisfiable in a total RA model according to D-semantics, then φ is satisfiable in a total RA model $\mathcal{M}$ with $|\mathcal{M}| \leq |\varphi|^{d(\varphi)}$.

Proof. By strong induction on $d(\varphi)$. Since φ is satisfiable iff $\neg\varphi$ is falsifiable, consider the latter. By Proposition 2.2, $\neg\varphi$ is equivalent to a conjunction of T-unpacked formulas of the form $\chi_{n,m}$, which is falsifiable iff one of its conjuncts $\chi_{n,m}$ is falsifiable. By Lemmas 2.2 - 2.4, if $\chi_{n,m}$ is falsifiable, then it is falsifiable in a model $\mathcal{M}$ that combines at most k other models (and one root world), where k is the number of top-level K operators in $\chi_{n,m}$, which is bounded by $|\varphi|$. Each of the these models is selected as a model of a formula of lesser modal depth than $\chi_{n,m}$, so by the inductive hypothesis we can assume that each is of size at most $|\varphi|^{d(\varphi)-1}$. Hence $|\mathcal{M}| \leq |\varphi| \times |\varphi|^{d(\varphi)-1} = |\varphi|^{d(\varphi)}$. □

Corollary 2.3 (Complexity of Satisfiability)**.**

1. The problem of deciding whether an epistemic formula is satisfiable in the class of total RA models according to D-semantics is in PSPACE;

2. For any k, the problem of deciding whether an epistemic formula φ with $d(\varphi) \leq k$ is satisfiable in the class of total RA models according to D-semantics is NP-complete.

Proof. (Sketch) For part 1, given PSPACE = NPSPACE (see Papadimitriou 1994, §7.3), it suffices to give a non-deterministic algorithm using polynomial space. By the previous results (including Prop. 2.2), if φ is satisfiable, then it is satisfiable in a model that can be inductively constructed as in the proofs of Lemmas 2.2, 2.3, and 2.4. We want an algorithm to non-deterministically guess such a model. However, since the size of the model may be exponential in $|\varphi|$, we cannot necessarily

store the entire model in memory using only polynomial space. Instead, we nondeterministically guess the submodels that are combined in the inductive construction, taking advantage of the following fact from the proof of Lemma 2.2. Once we have computed the truth values at $\mathcal{N}, v$ (or $\mathcal{O}, u$) of all subformulas of φ (up to some modal depth, depending on the stage of the construction), we can label v with the true subformulas and then erase the rest of $\mathcal{N}$ from memory (and similarly for $\mathcal{O}, u$). The other worlds in $\mathcal{N}$ will not be in the field of $\preceq_x$ for any world x at which we need to compute truth values at any later stage of the construction, so it is not necessary to access those worlds in order to compute later truth values. Given this space-saving method, we only need to use polynomial space at any given stage of the algorithm. I leave the details of the algorithm to the reader.[38]

For part 2, NP-hardness is immediate, since for $k = 0$ we have all formulas of propositional logic. For membership in NP, if φ is satisfiable and $d(\varphi) \leq k$, then by Corollary 2.2, φ satisfiable in a model $\mathcal{M}$ with $|\mathcal{M}| \leq |\varphi|^k$. We can nondeterministically guess such a model, and it is easy to check that evaluating φ in $\mathcal{M}$ is in polynomial time given that $\mathcal{M}$ is polynomial-sized. □

As explained in Remark 2.9, Corollary 2.3.1 accords with results of Vardi [1989]. Corollary 2.3.2 accords with results of Halpern [1995] on the effect of bounding modal depth on the complexity of satisfiability for modal logics.

2.6.3 Completeness for All RA Models

Next we prove the 'only if' direction of Theorem 2.1.3. In the process we prove the separation property for D-semantics over all RA models noted in Proposition 2.1. Interestingly, dropping totality makes things simpler.

Claim 2.2. If neither (a) nor (d) holds for a T-unpacked $\chi_{n,m}$, then there is a pointed RA model $\mathcal{M}, w$ such that $\mathcal{M}, w \not\models_d \chi_{n,m}$.

Proof. If $m \leq 1$, (d) is the same as (c), covered in §2.6.2. So suppose $m > 1$. By Lemma 2.4 and the $m = 1$ case of the inductive proof of Lemma 2.3, if neither

[38] Cf. Theorem 4.2 of Friedman and Halpern 1994 for a proof that the complexity of satisfiability for formulas of conditional logic in similar preorder structures is in PSPACE.

(a) nor (d) holds for $\chi_{n,m}$, then for all $1 \leq j \leq m$, there is a linear RA model $\mathcal{M}_j = \langle W_j, \rightarrow_j, \preceq^j, V_j \rangle$ with point $w_j \in W_j$ such that

$$\mathcal{M}_j, w_j \not\models_d \varphi_0 \wedge K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi_j. \tag{2.36}$$

Recall that $\mathcal{M}_j$ is constructed in such a way that for all $v \in W_j^- = W_j \setminus \{w_j\}$, w_j is not in the field of $\preceq_v^j$. Without loss of generality, assume that for all $j, k \leq m$, $W_j \cap W_k = \emptyset$. Construct $\mathcal{M} = \langle W, \rightarrow, \preceq, V \rangle$ as follows, by first taking the disjoint union of all of the $\mathcal{M}_j$, then "merging" all of the w_j into a single new world w (with the same valuation as some w_k), so that the linear models $\mathcal{M}_j$ are linked to w like spokes to the hub of a wheel (recall Fig. 2.3):

$$W = \{w\} \cup \bigcup_{j \leq m} W_j^-; \text{ for all } j \leq m \text{ and } v \in W_j^-, \preceq_v = \preceq_v^j;$$

$$\preceq_w = \{\langle w, v \rangle \mid v = w \text{ or } \exists\, j \leq m \colon w_j \preceq_{w_j}^j v\} \cup \bigcup_{j \leq m} (\preceq_{w_j}^j \cap (W_j^- \times W_j^-));$$

$$\rightarrow = \{\langle w, v \rangle \mid v = w \text{ or } \exists\, j \leq m \colon w_j \rightarrow_j v\} \cup \bigcup_{j \leq m} (\rightarrow_j \cap (W_j^- \times W_j^-));$$

$$V(p) = \begin{cases} \bigcup\limits_{j \leq m} (V_j(p) \cap W_j^-) \cup \{w\} & \text{if } w_1 \in V_1(p); \\ \bigcup\limits_{j \leq m} (V_j(p) \cap W_j^-) & \text{if } w_1 \notin V_1(p). \end{cases}$$

It is easy to verify that for all formulas ξ, $j \leq m$, and $v \in W_j^-$,

$$\mathcal{M}, v \vDash_d \xi \text{ iff } \mathcal{M}_j, v \vDash_d \xi. \tag{2.37}$$

It follows from the construction of $\mathcal{M}$ and (2.37) that for all $j \leq m$,

$$\mathrm{Min}_{\preceq_{w_j}^j}(\overline{[\![\psi_j]\!]}^{\mathcal{M}_j}) \cap \rightarrow_j(w) \subseteq \mathrm{Min}_{\preceq_w}(\overline{[\![\psi_j]\!]}^{\mathcal{M}}) \cap \rightarrow(w). \tag{2.38}$$

For all $j \leq m$, given $\mathcal{M}_j, w_j \not\models_d K\psi_j$ by assumption, the left side of (2.38) is nonempty, so the right side is nonempty. Hence by the truth definition,

$$\mathcal{M}, w \not\models_d K\psi_1 \vee \cdots \vee K\psi_m. \tag{2.39}$$

By our initial assumption, for all $j \leq m$,

$$\bigcup_{i \leq n} \mathrm{Min}_{\preceq^j_{w_j}}(\overline{[\![\varphi_i]\!]}^{\mathcal{M}_j}) \cap \rightarrow^j(w) = \emptyset. \tag{2.40}$$

We prove by induction that for $1 \leq i \leq n$,

$$\mathrm{Min}_{\preceq_w}(\overline{[\![\varphi_i]\!]}^{\mathcal{M}}) \cap \rightarrow(w) = \emptyset. \tag{2.41}$$

Base case. Given $\mathcal{M}_1, w_1 \vDash \varphi_0$ and the fact that w has the same valuation under V as w_1 under V_1, we have $\mathcal{M}, w \vDash \varphi_0$. Together with (2.39), this implies $\mathcal{M}, w \nvDash_d \chi_{0,m}$. Since $\chi_{1,m}$ is T-unpacked, it follows by Definition 2.9 that $\mathcal{M}, w \vDash_d \varphi_1$, in which case $w \notin \mathrm{Min}_{\preceq_w}(\overline{[\![\varphi_1]\!]}^{\mathcal{M}})$. By construction of $\mathcal{M}$, together (2.40), (2.37), and $w \notin \mathrm{Min}_{\preceq_w}(\overline{[\![\varphi_1]\!]}^{\mathcal{M}})$ imply (2.41) for $i = 1$.

Inductive step. Assume (2.41) for all $k \leq i$ $(i < n)$, so $\mathcal{M}, w \vDash_d K\varphi_1 \wedge \cdots \wedge K\varphi_i$, which with (2.39) gives $\mathcal{M}, w \nvDash_d \chi_{i,m}$. Then since $\chi_{i+1,m}$ is T-unpacked, $\mathcal{M}, w \vDash_d \varphi_{i+1}$, so by reasoning as in the base case, (2.41) holds for $i + 1$.

Since (2.41) holds for $1 \leq i \leq n$, by the truth definition we have $\mathcal{M}, w \vDash_d K\varphi_1 \wedge \cdots \wedge K\varphi_n$, which with $\mathcal{M}, w \vDash \varphi_0$ and (2.39) implies $\mathcal{M}, w \nvDash_d \chi_{n,m}$. □

A remark analogous to Remark 2.8 applies to the above construction: if each $\rightarrow_j$ is an equivalence relation and we extend $\rightarrow$ to the minimal equivalence relation $\rightarrow^+ \supseteq \rightarrow$, then the resulting model will still falsify $\chi_{n,m}$. Hence Theorem 2.1.3 holds for the class of RA models with equivalence relations (and with the universal field property by Prop. 2.3). Finally, arguments similar to those of Corollaries 2.2 - 2.3 show the finite model property and PSPACE satisfiability without the assumption of totality (see Remark 2.9).

2.6.4 Completeness for CB Models

Finally, for the 'only if' direction of Theorem 2.1.4, there are two ways to try to falsify some $\chi_{n,m}$. For H/N-semantics, we can first construct an RA countermodel for $\chi_{n,m}$ under D-semantics, as in §2.6.2, and then transform it into a CB countermodel

for $\chi_{n,m}$ under H/N-semantics, as shown in §2.7 below. Alternatively, we can first construct a CB countermodel under S/H-semantics and then transform it into a CB countermodel under H/N-semantics as in §2.7. Here we will take the latter route. By Proposition 2.5 below, for the 'only if' direction of Theorem 2.1.4 it suffices to prove the following.

Claim 2.3. If neither (a) nor (d) holds for a flat, T-unpacked $\chi_{n,m}$, then there is a pointed CB model $\mathcal{M}, w$ such that $\mathcal{M}, w \nvDash_{h,s} \chi_{n,m}$.

We begin with some notation used in the proof and in later sections.

Notation 2.5 (Relational Image)**.** Given a CB model $\mathcal{M} = \langle W, D, \leqslant, V\rangle$, the image of $\{w\}$ under the relation D is $D(w) = \{v \in W \mid wDv\}$.

Hence $D(w)$ is the set of doxastically accessible worlds for the agent in w.

Let us now prove the claim.

Proof. For any positive integer z, let $P_z = \{1, \ldots, z\}$. For all $k \in P_m$, let $S_k = \{i \in P_n \mid \vDash \psi_k \to \varphi_i\}$, and $T = \{t \in P_m \mid S_t = P_n\}$. Since (d) does not hold for $\chi_{n,m}$, it follows that

$$\nvDash \bigwedge_{i \in S_k} \varphi_i \to \psi_k. \tag{2.42}$$

Construct $\mathcal{M} = \langle W, D, \leqslant, V\rangle$ as follows (see Fig. 2.6):

$W = \{w\} \cup \{x_t \mid t \in T\} \cup \{v_k, u^k_j \mid k \in P_m \setminus T \text{ and } j \in P_n \setminus S_k\}$;

D is the union of $\{\langle w, w\rangle\}$, $\{\langle w, x_t\rangle, \langle x_t, x_t\rangle \mid t \in T\}$, and

$$\{\langle v_k, u^k_j\rangle, \langle u^k_j, u^k_j\rangle \mid k \in P_m \setminus T \text{ and } j \in P_n \setminus S_k\};$$

$\leqslant_w = \{\langle w, w\rangle\} \cup \{\langle w, v_k\rangle, \langle v_k, w\rangle, \langle v_k, v_k\rangle \mid k \in P_m\}$;[39]

[39]The x_t and u^k_j worlds are not in the field of $\leqslant_w$. For a universal field (and total relation), the proof works with minor additions if we take the union of $\leqslant_w$ as defined above with

$$\{\langle w, x_t\rangle, \langle w, u^k_j\rangle, \langle v_k, x_t\rangle, \langle v_k, u^k_j\rangle, \langle x_t, x_t\rangle, \langle x_t, u^k_j\rangle, \langle u^k_j, u^k_j\rangle \mid t \in T,\, k \in P_m \setminus T,\, j \in P_n \setminus S_k\}.$$

For $y \in W \setminus \{w\}$, $\leqslant_y$ is any relation as in Definition 2.2.3;

V is any valuation function on W such that $\mathcal{M}, w \vDash \varphi_0$ and

- for all $t \in T$, $\mathcal{M}, x_t \vDash \bigwedge_{i \in P_n} \varphi_i \wedge \neg\psi_t$;
- for all $k \in P_m \setminus T$, $\mathcal{M}, v_k \vDash \bigwedge_{i \in S_k} \varphi_i \wedge \neg\psi_k$;
- for all $k \in P_m \setminus T$ and $j \in P_n \setminus S_k$, $\mathcal{M}, u_j^k \vDash \neg\varphi_j \wedge \psi_k$.

Such a valuation V exists by the assumption that (a) does not hold for $\chi_{n,m}$, together with (2.42) and the definitions of T and S_k.

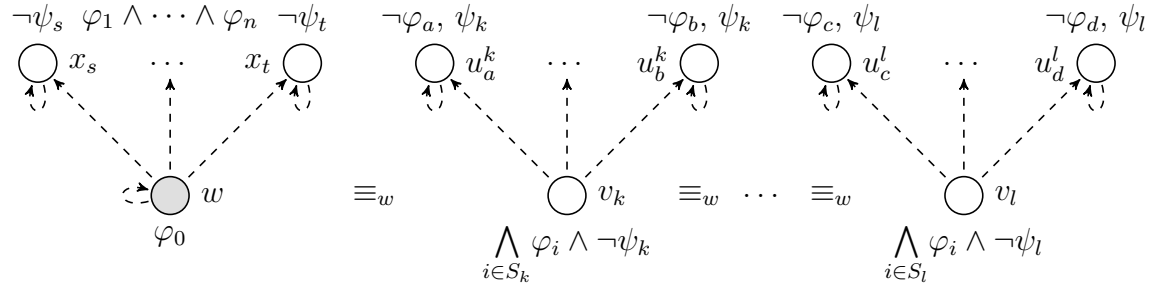

Figure 2.6: countermodel for $\chi_{n,m}$ in H/S-semantics

Since $\chi_{n,m}$ is flat and T-unpacked, $\mathcal{M}, w \vDash \varphi_0$ implies $\mathcal{M}, w \vDash \varphi_1 \wedge \cdots \wedge \varphi_n$. Then since $D(w) = \{w\} \cup \{x_t \mid t \in T\}$ and $\mathcal{M}, x_t \vDash \varphi_1 \wedge \cdots \wedge \varphi_n$ for all $t \in T$,

$$\mathcal{M}, w \vDash \bigwedge_{i \in P_n} (B\varphi_i \wedge \varphi_i). \tag{2.43}$$

For all $k \in P_m \setminus T$, we have

$$\mathcal{M}, v_k \nvDash \bigvee_{j \in P_n \setminus S_k} B\varphi_j \tag{2.44}$$

given $v_k D u_j^k$ and $\mathcal{M}, u_j^k \nvDash \varphi_j$, and

$$\mathcal{M}, v_k \vDash \bigwedge_{i \in S_k} \varphi_i \tag{2.45}$$

by definition of V. It follows from (2.44) and (3.3.2) that for all $k \in P_m \setminus T$,

$$\mathcal{M}, v_k \vDash \bigwedge_{i \in P_n} (B\varphi_i \to \varphi_i). \tag{2.46}$$

By construction of $\mathcal{M}$, (2.43) and (2.46) together imply that for all $y \in W_w$,

$$\mathcal{M}, y \vDash \bigwedge_{i \in P_n} (B\varphi_i \to \varphi_i). \tag{2.47}$$

Together (2.43) and (2.47) imply $\mathcal{M}, w \vDash_{h,s} K\varphi_i$ for all $i \in P_n$ by the truth definitions (Def. 2.7). Now let us check that $\mathcal{M}, w \nvDash_{h,s} K\psi_i$ for all $i \in P_m$. On the one hand, for all $t \in T$, given wDx_t and $\mathcal{M}, x_t \nvDash \psi_t$, we have $\mathcal{M}, w \nvDash B\psi_t$ and hence $\mathcal{M}, w \nvDash_{h,s} K\psi_t$. On the other hand, for all $k \in P_m \setminus T$, given $D(v_k) = \{u_k^j \mid j \in P_n \setminus S_k\}$ and $\mathcal{M}, u_j^k \vDash \psi_k$, we have $\mathcal{M}, v_k \vDash B\psi_k$; but then since $\mathcal{M}, v_k \nvDash \psi_k$ and $v_k \in \mathrm{Min}_{\leqslant_w}(W)$, it follows that $\mathcal{M}, w \nvDash_{h,s} K\psi_k$. Together with $\mathcal{M}, w \vDash \varphi_0$, the previous facts imply $\mathcal{M}, w \nvDash_{h,s} \chi_{n,m}$. □

We leave the extension of the 'only if' direction of Theorem 2.1.4 to the full epistemic language for other work (see Problem 2.1). Facts 2.9.4, 2.9.5, and 2.11.1 show that for the full language, this direction must be modified. Yet for our purposes here, the above proof already helps to reveal the sources of closure failure in H/S-semantics and in N-semantics by Proposition 2.5 below.

2.6.5 The Sources of Closure Failure

The results of §2.6.2 - 2.6.4 allow us to clearly identify the sources of closure failure in D/H/N/S-semantics. In D-semantics, the source of closure failure is the orderings—if we collapse the orderings, then D- is equivalent to L-semantics (see Observation 2.3) and closure failures disappear. By Proposition 2.4 below, the orderings are also a source of closure failure in H/N-semantics. However, the proof in §2.6.4 shows that there is another source of closure failure in H/N/S-semantics: the interpretation of ruling out in terms *belief*, as in the quote from Heller in §2.4. This is the sole source of closure failure in S-semantics, the odd member of the D/H/N/S-family that does

not use the orderings beyond $\mathrm{Min}_{\leqslant_w}(W)$ (recall Observation 2.1). Given this source of closure failure, even if we collapse the orderings, in which case H- is equivalent to S-semantics (see Prop. 2.6), closure failure persists. We will return to this point in §2.10.

2.7 Relating RA and CB Models

The discussion in §2.6.4 - 2.6.5 appealed to claims about the relations between D/H/N/S-semantics. In this short section, we prove these claims. Readers eager to see how the results of §2.6 lead to complete deductive systems for the RA and subjunctivist theories should skip ahead to §2.8 and return here later.

One way to see how the RA and subjunctivist theories are related is by transforming models viewed from the perspective of one theory into models that are equivalent, with respect to what can be expressed in our language, when viewed from the perspective of another theory. This also shows that any closure principle that fails for the first theory also fails for the second.

We first see how to transform any RA model viewed from the perspective of D-semantics into a CB model that is equivalent, with respect to the flat fragment of the epistemic language, when viewed from the perspective of H-semantics. The transformation is intuitive: if, in the RA model, a possibility v is *eliminated* by the agent in w, then we construct the CB model such that if the agent were in situation v instead of w, the agent would *notice*, i.e., would correctly believe that the true situation is v rather than w;[40] but if, in the RA model, v is *uneliminated* by the agent in w, then we construct the CB model such that if the agent were in situation v instead of w, the agent would *not* notice, i.e., would incorrectly believe that the true situation is w rather than v. (The CB model in Fig. 2.2 is obtained from the RA model in Fig. 2.1 in this way.) Then the agent has eliminated the relevant alternatives to a flat φ at w in the RA model iff the agent *sensitively* believes φ at w in the CB

[40]In fact, we only need something weaker, namely, that it would be *compatible* with what the agent believes that the true situation is v, i.e., vDv. In the $w \not\rightarrow v$ case of the definition of D in the proof of Proposition 2.4, we only need that $v \in D(v)$ for the proof to work.

model.

Proposition 2.4 (D-to-H Transform)**.** For any RA model $\mathcal{M} = \langle W, \rightarrowtriangle, \preceq, V\rangle$ with $w \in W$, there is a CB model $\mathcal{N} = \langle W, D, \leqslant, V\rangle$ such that for all flat epistemic formulas φ,

$$\mathcal{M}, w \vDash_d \varphi \text{ iff } \mathcal{N}, w \vDash_h \varphi.$$

Proof. Construct $\mathcal{N}$ from $\mathcal{M}$ as follows. Let W and V in $\mathcal{N}$ be the same as in $\mathcal{M}$; let $\leqslant$ in $\mathcal{N}$ be the same as $\preceq$ in $\mathcal{M}$; construct D in $\mathcal{N}$ from $\rightarrowtriangle$ in $\mathcal{M}$ as follows, where w is the fixed world in the lemma (recall Notation 2.5):

$$\forall v \in W\colon D(v) = \begin{cases} \{w\} \text{ if } w \rightarrowtriangle v; \\ \{v\} \text{ if } w \not\rightarrowtriangle v. \end{cases} \tag{2.48}$$

To prove the 'iff' by induction on φ, the base case is immediate and the boolean cases routine. Suppose φ is of the form $K\psi$. Since φ is flat, ψ is propositional. Given that V is the same in $\mathcal{N}$ as in $\mathcal{M}$, for all $v \in W$, $\mathcal{M}, v \vDash_d \psi$ iff $\mathcal{N}, v \vDash_h \psi$. Hence if $\mathcal{M}, w \nvDash_d \psi$, then $\mathcal{M}, w \nvDash_d K\psi$ and $\mathcal{N}, w \nvDash_h K\psi$ by Facts 2.1 and 2.4. Suppose $\mathcal{M}, w \vDash_d \psi$. Since $w \rightarrowtriangle w$, we have $D(w) = \{w\}$ by construction of $\mathcal{N}$, so $\mathcal{N}, w \vDash_h B\psi$ given $\mathcal{N}, w \vDash_h \psi$. It only remains to show that $\mathcal{M}, w \vDash_d K\psi$ iff the sensitivity condition (Def. 2.7) for $K\psi$ is satisfied at $\mathcal{N}, w$. This is easily seen to be a consequence of the following, given by the construction of $\mathcal{N}$:

$$\mathrm{Min}_{\preceq_w}(\overline{[\![\psi]\!]}_d^{\mathcal{M}}) = \mathrm{Min}_{\leqslant_w}(\overline{[\![\psi]\!]}_h^{\mathcal{N}}); \tag{2.49}$$

$$\forall u \in \mathrm{Min}_{\preceq_w}(\overline{[\![\psi]\!]}_d^{\mathcal{M}})\colon w \rightarrowtriangle u \text{ iff } \mathcal{N}, u \vDash_h B\psi. \tag{2.50}$$

The left-to-right direction of the biconditional in (2.50) follows from the fact that if $w \rightarrowtriangle u$, then $D(u) = \{w\}$, and $\mathcal{N}, w \vDash_h \psi$. For the right-to-left direction, if $w \not\rightarrowtriangle u$, then $D(u) = \{u\}$, in which case $\mathcal{N}, u \nvDash_h B\psi$ given $\mathcal{N}, u \nvDash_h \psi$. □

The transformation above does not always preserve all non-flat epistemic formulas, and by Fact 2.9.4, no transformation does so. However, since the flat fragment of the language suffices to express all principles of closure with respect to propositional

logic, Proposition 2.4 has the notable corollary that all such closure principles that fail in D-semantics also fail in H-semantics.

Next we transform CB models viewed from the perspective of H-semantics into CB models that are equivalent, with respect to the epistemic-doxastic language, when viewed from the perspective of N-semantics. (Fact 2.9 in §2.9 shows that there is no such general transformation in the N-to-H direction.) To do so, we make the models *centered*, which (as noted in Observation 2.1) trivializes the adherence condition that separates N- from H-semantics.

Proposition 2.5 (H-to-N Transform)**.** For any CB model $\mathcal{N} = \langle W, D, \leqslant, V\rangle$, there is a CB model $\mathcal{N}' = \langle W, D, \leqslant', V\rangle$ such that for all $w \in W$ and all epistemic-doxastic formulas φ,

$$\mathcal{N}, w \vDash_h \varphi \text{ iff } \mathcal{N}', w \vDash_n \varphi.$$

Proof. Construct $\mathcal{N}'$ from $\mathcal{N}$ as follows. Let W, D, and V in $\mathcal{N}'$ be the same as in $\mathcal{N}$. For all $w \in W$, construct $\leqslant'_w$ from $\leqslant_w$ by making w strictly minimal in $\leqslant'_w$, but changing nothing else:

$$u \leqslant'_w v \text{ iff } \begin{cases} v \neq w \text{ and } u \leqslant_w v, \text{ or} \\ u = w. \end{cases} \tag{2.51}$$

To prove the proposition by induction on φ, the base case is immediate and the boolean and belief cases routine. Suppose φ is $K\psi$ and $[\![\psi]\!]_h^{\mathcal{N}} = [\![\psi]\!]_n^{\mathcal{N}'}$. If $\mathcal{N}, w \nvDash_h \psi$, then $\mathcal{N}, w \nvDash_h K\psi$ and $\mathcal{N}', w \nvDash_n K\psi$ by Fact 2.4. If $\mathcal{N}, w \vDash_h \psi$ and hence $\mathcal{N}', w \vDash_n \psi$, then by construction of $\leqslant'_w$ and the inductive hypothesis,

$$\mathrm{Min}_{\leqslant_w}(\overline{[\![\psi]\!]}_h^{\mathcal{N}}) = \mathrm{Min}_{\leqslant'_w}(\overline{[\![\psi]\!]}_n^{\mathcal{N}'}). \tag{2.52}$$

Since D is the same in $\mathcal{N}$ as in $\mathcal{N}'$, (2.52) implies that the belief and sensitivity conditions for $K\psi$ are satisfied at $\mathcal{N}, w$ iff they are satisfied at $\mathcal{N}', w$. If the belief condition is satisfied, then $\mathrm{Min}_{\leqslant'_w}([\![B\psi]\!]_n^{\mathcal{N}'}) = \{w\}$ by construction of $\leqslant'_w$, so the adherence condition (Def. 2.7) is automatically satisfied at $\mathcal{N}', w$. Hence the belief and sensitivity conditions for $K\psi$ are satisfied at $\mathcal{N}, w$ iff the belief, sensitivity, and

adherence conditions are satisfied for $K\psi$ at $\mathcal{N}', w$.[41] □

Our last transformation takes us from models viewed from the perspective of S-semantics to equivalent models viewed from the perspective of H-semantics—and hence N-semantics by Proposition 2.5. (Fact 2.8 in §2.9 shows that there can be no such general transformation in the H-to-S direction.) The idea of the transformation is that safety is the $\exists\forall$ condition (as in §2.4) obtained by restricting the scope of sensitivity to a fixed set of worlds, $\mathrm{Min}_{\leqslant_w}(W)$.

Proposition 2.6 (S-to-H Transform)**.** For any CB model $\mathcal{N} = \langle W, D, \leqslant, V\rangle$, there is a CB model $\mathcal{N}' = \langle W, D, \leqslant', V\rangle$ such that for all $w \in W$ and all epistemic-doxastic formulas φ,

$$\mathcal{N}, w \vDash_s \varphi \text{ iff } \mathcal{N}', w \vDash_h \varphi.$$

Proof. Construct $\mathcal{N}'$ from $\mathcal{N}$ as follows. Let W, D, and V in $\mathcal{N}'$ be the same as in $\mathcal{N}$. For all $w \in W$, construct $\leqslant'_w$ from $\leqslant_w$ by taking $\mathrm{Min}_{\leqslant_w}(W)$ to be the field of $\leqslant'_w$ and setting $u \leqslant'_w v$ for all u and v in the field. It is straightforward to check that $\mathcal{N}$ and $\mathcal{N}'$ are equivalent with respect to the safety condition and that in $\mathcal{N}'$ the safety and sensitivity conditions become equivalent.[42] □

Although I have introduced the propositions above for the purpose of relating the (in)valid closure principles of one theory to those of another, by transforming countermodels of one kind into countermodels of another, the interest of this style of analysis is not just in transferring principles for reasoning about knowledge between theories; the interest is also in highlighting the structural relations between different pictures of what knowledge is. In Chapter 3, we will continue our model-theoretic analysis to illuminate these pictures.

[41] It is easy to see that even if we forbid centered models, Proposition 2.5 will still hold. For we can allow any number of worlds in $\mathrm{Min}_{\leqslant'_w}(W)$, provided they do not witness a violation of the adherence condition at w for any φ for which we want $\mathcal{N}, w \vDash_n K\varphi$.

[42] It is easy to see that even if we require $W_w \setminus \mathrm{Min}_{\leqslant'_w}(W) \neq \emptyset$, Proposition 2.5 will still hold. For we can allow any number of worlds in $W_w \setminus \mathrm{Min}_{\leqslant'_w}(W)$, provided they do not witness a violation of the sensitivity condition at w for any φ for which we want $\mathcal{N}, w \vDash_h K\varphi$.

2.8 Deductive Systems

From Theorem 2.1 we obtain complete deductive systems for reasoning about knowledge according to the RA, tracking, and safety theories. Table 2.1 lists all of the needed schemas and rules, using the nomenclature of Chellas [1980] (except for X, RAT, and RA, which are new). **E** is the weakest of the *classical* modal systems with PL, MP, and RE. $\mathbf{ES}_1 \dots \mathbf{S}_n$ is the extension of **E** with every instance of schemas $\mathrm{S}_1 \dots \mathrm{S}_n$. **EMCN** is familiar as the weakest normal modal system **K**, equivalently characterized in terms of PL, MP, the K schema, and the necessitation rule for K (even more simply, by PL, MP, and RK).

Corollary 2.4 (Soundness and Completeness)**.**

1. The system **KT** (equivalently, **ET** plus the RK rule) is sound and complete for C/L-semantics over RA models.

2. (The Logic of Ranked Relevant Alternatives) The system **ECNTX** (equivalently, **ET** plus the RAT rule) is sound and complete for D-semantics over total RA models.

3. The system **ECNT** (equivalently, **ET** plus the RA rule) is sound and complete for D-semantics over RA models.

4. **ECNT** is sound (with respect to the full epistemic language) and complete (with respect to the flat fragment) for H/N/S-semantics over CB models.[43]

The proof of Corollary 2.4 is similar to the alternative completeness proof discussed by van Benthem [2010, §4.3] for the system **K**.[44]

[43]Corollary 2.4.4 gives an answer, for the flat fragment, to the question posed by van Benthem [2010, 153] of what is the epistemic logic of Nozick's notion of knowledge.

[44]The usual canonical model approach used for **K** and other normal modal logics seems more difficult to apply to RA and CB models, since we must use maximally consistent sets of formulas in the epistemic language only (cf. Remark 2.10) to guide the construction of both the orderings $\preceq_w$ (resp. $\leqslant_w$) and relation $\rightarrowtriangle$ (resp. D), which must be appropriately related to one another for the truth lemma to hold. In this situation, our alternative approach performs well.

PL. all tautologies		MP. $\dfrac{\varphi \to \psi \quad \varphi}{\psi}$
T. $K\varphi \to \varphi$	N. $K\top$	RE. $\dfrac{\varphi \leftrightarrow \psi}{K\varphi \leftrightarrow K\psi}$
M. $K(\varphi \wedge \psi) \to K\varphi \wedge K\psi$		RK. $\dfrac{\varphi_1 \wedge \cdots \wedge \varphi_n \to \psi}{K\varphi_1 \wedge \cdots \wedge K\varphi_n \to K\psi}$ $(n \geq 0)$
X. $K(\varphi \wedge \psi) \to K\varphi \vee K\psi$		RAT. $\dfrac{\varphi_1 \wedge \cdots \wedge \varphi_n \leftrightarrow \psi_1 \wedge \cdots \wedge \psi_m}{K\varphi_1 \wedge \cdots \wedge K\varphi_n \to K\psi_1 \vee \cdots \vee K\psi_m}$ $(n \geq 0, m \geq 1)$
C. $K\varphi \wedge K\psi \to K(\varphi \wedge \psi)$		RA. $\dfrac{\varphi_1 \wedge \cdots \wedge \varphi_n \leftrightarrow \psi}{K\varphi_1 \wedge \cdots \wedge K\varphi_n \to K\psi}$ $(n \geq 0)$

Table 2.1: axiom schemas and rules

Proof. We only give the proof for part 2, since the proofs for the others are similar. Soundness follows from Theorem 2.1.2. For completeness, we first prove by strong induction on the modal depth $d(\varphi)$ of φ (Def. 2.1) that if φ is D-valid over total RA models, then φ is provable in the system combining **ET** and the RAT rule. If $d(\varphi) = 0$, then the claim is immediate, since our deductive system includes propositional logic. Suppose $d(\varphi) = n + 1$. By the proof of Proposition 2.2, using PL, MP, T, and RE (which is a derived rule given RAT, PL, and MP), we can prove that φ is equivalent to a conjunction φ', each of whose conjuncts is a *T-unpacked* formula (Def. 2.9) of the form

$$\varphi_0 \wedge K\varphi_1 \wedge \cdots \wedge K\varphi_n \to K\psi_1 \vee \cdots \vee K\psi_m. \tag{2.53}$$

The conjunction φ' is valid iff each conjunct of the form of (2.53) is valid. By Theorem 2.1.2, (2.53) is valid iff either condition (a) or condition (c) of Theorem 2.1.2 holds. *Case 1*: (a) holds, so $\varphi_0 \to \bot$ is valid. By the inductive hypothesis, we can derive $\varphi_0 \to \bot$, from which we derive (2.53) using PL and MP. *Case 2*: (c) holds, so for some $\Phi \subseteq \{\varphi_1, \ldots, \varphi_n\}$ and nonempty $\Psi \subseteq \{\psi_1, \ldots, \psi_m\}$,

$$\bigwedge_{\varphi \in \Phi} \varphi \leftrightarrow \bigwedge_{\psi \in \Psi} \psi \tag{2.54}$$

is valid. Since (2.54) is of modal depth less than $n+1$, by the inductive hypothesis it is provable. From (2.54), we can derive

$$\bigwedge_{\varphi\in\Phi} K\varphi \to \bigvee_{\psi\in\Psi} K\psi \tag{2.55}$$

using the RAT rule, from which we can derive (2.53) using PL and MP. Having derived each conjunct of φ' in one of these ways, we can use PL and MP to derive the conjunction itself, which by assumption is provably equivalent to φ.

Next we show by induction on the length of proofs that any proof in the system combining **ET** and RAT can be transformed into an **ECNTX** proof of the same theorem. Suppose that in the first proof, $\varphi_1 \wedge \cdots \wedge \varphi_n \leftrightarrow \psi_1 \wedge \cdots \wedge \psi_m$ has been derived, to which the RAT rule is applied. In the second proof, if $n > 0$, we first derive $K\varphi_1 \wedge \cdots \wedge K\varphi_n \to K(\varphi_1 \wedge \cdots \wedge \varphi_n)$ using C repeatedly (with PL and MP); next, we derive $K(\varphi_1 \wedge \cdots \wedge \varphi_n) \leftrightarrow K(\psi_1 \wedge \cdots \wedge \psi_m)$ by applying the RE rule to $\varphi_1 \wedge \cdots \wedge \varphi_n \leftrightarrow \psi_1 \wedge \cdots \wedge \psi_m$; we then derive $K(\psi_1 \wedge \cdots \wedge \psi_m) \to K\psi_1 \vee \cdots \vee K\psi_m$ using X repeatedly (with PL and MP); finally, we derive $K\varphi_1 \wedge \cdots \wedge K\varphi_n \to K\psi_1 \vee \cdots \vee K\psi_m$ using PL, MP, and earlier steps. If $n = 0$,[45] we first derive $K\top$ using N, then derive $K\top \leftrightarrow K(\psi_1 \wedge \cdots \wedge \psi_m)$ by applying the RE rule to $\top \leftrightarrow \psi_1 \wedge \cdots \wedge \psi_m$, then derive the conclusion of the RAT application using X, PL, and MP. □

For reasons suggested in §2.1, I do not consider the systems of Corollary 2.4.2-.4 to be plausible as *epistemic* logics, and therefore I do not consider the basic theories they are based on to be satisfactory theories of knowledge. Nonetheless, we may wish to reason directly about whether one has ruled out the relevant alternatives, whether one's beliefs are sensitive to the truth, etc., and Corollary 2.4 gives principles for these notions. Simply replace the K symbol by a neutral $\Box$ and the newly identified logic **ECNTX**, which I dub *the logic of ranked relevant alternatives*, is of significant independent interest.

With these qualifications in mind, I will make another negative point concerning knowledge. It is straightforward to derive the K axiom, the star of the epistemic

[45] If $n = 0$, we can take the left side of the premise/conclusion of RAT to be $\top$, or we can simply take the premise to be $\psi_1 \vee \cdots \vee \psi_m$ and the conclusion to be $K\psi_1 \vee \cdots \vee K\psi_m$.

closure debate with its leading role in skeptical arguments, from M, C, RE, and propositional logic. Hence in order to avoid K one must give up one of the latter principles. (For RE, recall that we are considering ideally astute logicians as in §2.1.) What is so strange about subjunctivist-flavored theories is that they validate C but not M, which seems to get things backwards. Hawthorne [2004a, §4.6, §1.6] discusses some of the problems and puzzles, related to the Lottery and Preface Paradoxes [Kyburg, 1961, Makinson, 1965], to which C leads (also see Goldman 1975). M seems rather harmless by comparison (cf. Williamson 2000, §12.2). Interestingly, C also leads to computational difficulties.

Remark 2.9 (NP vs. PSPACE)**.** Vardi [1989] proved a PSPACE upper bound for the complexity of the system **ECNT**,[46] in agreement with our conclusion in §2.6.3. (Together Corollaries 2.3 and 2.4.2 give a PSPACE upper bound for **ECNTX**.) Vardi also conjectured a PSPACE lower bound for **ECNT**. By contrast, he showed that for any subset of $\{\mathrm{T}, \mathrm{N}, \mathrm{M}\}$ added to **E**, complexity drops to NP-complete. Hence Vardi conjectured that the C axiom is the culprit behind the jump in complexity of epistemic logics from NP to PSPACE.[47] It appears that not only is C more problematic than M epistemologically, but also it makes reasoning about knowledge more computationally costly.[48]

[46]Here I mean either the problem of checking provability/validity or that of checking consistency/satisfiability, given that PSPACE is closed under complementation. When I refer to NP-completeness, I have in mind the consistency/satisfiability problem.

[47]In fact, Allen [2005] shows that adding any degree of conjunctive closure, however weak, to the classical modal logic **EMN** results in a jump from NP- to PSPACE-completeness. Adding the full strength of C is sufficient, but not necessary. As far as I know, lower bounds for the complexity of systems with C but without M have not yet been established.

[48]Whether such complexity facts have any philosophical significance seems to be an open question. As a cautionary example, one would not want to argue that it counts in favor of the plausibility of the 5 axiom, $\neg K\varphi \to K\neg K\varphi$, that while the complexity of **K** is PSPACE-complete, for any extension of **K5**, complexity drops to NP-complete [Halpern and Rêgo, 2007]. That being said, if we are forced to give up C for epistemological reasons, then its computational costliness in reasoning about knowledge may make us miss it less.

2.9 Higher-Order Knowledge

In this section, we briefly explore how the theories formalized in §2.4 - 2.5 differ with respect to knowledge about one's own knowledge and beliefs. The result is a hierarchical picture (Corollary 2.5) and an open problem for future research. First, we discuss a subtlety concerning higher-order RA knowledge. Second, we relate properties of higher-order subjunctivist knowledge to closure failures.

2.9.1 Higher-Order Knowledge and Relevant Alternatives

Theorem 2.1 and Corollary 2.4 show that no non-trivial principles of higher order knowledge, such as the controversial 4 axiom $K\varphi \to KK\varphi$ and 5 axiom $\neg K\varphi \to K\neg K\varphi$, are valid over RA models according to either L- or D-semantics. This is so even if we assume that the relation $\twoheadrightarrow$ in our RA models is an equivalence relation (see Remark 2.8), following Lewis [1996].

Example 2.3 (Failure of 4 Axiom)**.** For the model $\mathcal{M}$ in Fig. 2.7, in which $\twoheadrightarrow$ is an equivalence relation, observe that $\mathcal{M}, w_1 \not\models_{l,d} Kp \to KKp$. Since $\mathrm{Min}_{\preceq_{w_1}}(W) \cap \overline{[\![p]\!]} = \{w_2\}$ and $w_1 \not\twoheadrightarrow w_2$, we have $\mathcal{M}, w_1 \models_{l,d} Kp$. By contrast, since $w_4 \in \mathrm{Min}_{\preceq_{w_3}}(W) \cap \overline{[\![p]\!]}$ and $w_3 \twoheadrightarrow w_4$, we have $\mathcal{M}, w_3 \not\models_{l,d} Kp$. It follows that $w_3 \in \mathrm{Min}_{\preceq_{w_1}}(W) \cap \overline{[\![Kp]\!]}$, in which case $\mathcal{M}, w_1 \not\models_{l,d} KKp$ given $w_1 \twoheadrightarrow w_3$.

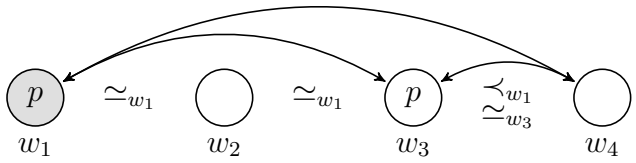

Figure 2.7: an RA countermodel for $Kp \to KKp$ in L/D-semantics (partially drawn, reflexive loops omitted)

According to Williamson [2001, 2009], "It is not always appreciated that ... since Lewis's accessibility relation is an equivalence relation, his account validates not only logical omniscience but the very strong epistemic logic S5" [2009, 23n16]. However,

Example 2.3 shows that this is not the case if we allow that comparative relevance, like comparative similarity, is possibility-relative, as seems reasonable for a Lewisian theory.[49] Other RA theorists are explicit that relevance depends on similarity of worlds (see, e.g., Heller 1989, 1999b), in which case the former should be world-relative since the latter is. For Williamson's point to hold, we would have to block the likes of Example 2.3 with an additional constraint on our models, such as the following.

Definition 2.11 (Absoluteness)**.** For an RA model $\mathcal{M} = \langle W, \rightarrow\!\!\!\!\triangleright, \preceq, V\rangle$, $\preceq$ is *locally* (resp. *globally*) *absolute* iff for all $w \in W$ and $v \in W_w$ (resp. for all $w, v \in W$), $\preceq_w = \preceq_v$ [Lewis, 1973, §6.1].

It is noteworthy that absoluteness leads to a collapse of comparative relevance.

Observation 2.2 (Absoluteness and Collapse)**.** Given condition 3b of Definition 2.2, if $\preceq$ is locally absolute, then for all $w \in W$ and $v \in W_w$,

$$\mathrm{Min}_{\preceq_w}(W) = W_w = \mathrm{Min}_{\preceq_v}(W) = W_v.$$

If $\preceq$ is globally absolute, then for all $w \in W$, $\mathrm{Min}_{\preceq_w}(W) = W$.

Lewis [1973, 99] rejected absoluteness for comparative similarity because it leads to such a collapse. We note that with the collapse of comparative relevance, the distinction between L- and D-semantics also collapses.

Observation 2.3 (Absoluteness and Collapse cont.)**.** Over locally absolute RA models, L- and D-semantics are equivalent.

[49] It follows from Lewis's [1996, 556f] *Rule of Resemblance* that if some $\neg p$-possibility w_2 "saliently resembles" w_1, which is relevant at w_1 by the Rule of Actuality, then w_2 is relevant at w_1, so you must rule out w_2 in order to know p in w_1. Lewis is explicit (555) that by 'actuality' he means the actuality of the subject of knowledge attribution. Hence if we consider your *counterpart* in some w_3, and some $\neg p$-possibility w_4 saliently resembles w_3, then your counterpart must rule out w_4 in order to know p in w_3. However, if salient resemblance is possibility-relative, as comparative similarity is for Lewis, then w_4 may not saliently resemble w_1, in which case you may not need to rule out w_4 in order to know p in w_1. (By Lewis's *Rule of Attention* (559), our attending to w_4 in this way may shift the context $\mathcal{C}$ to a context $\mathcal{C}'$ in which w_4 is relevant, but the foregoing points still apply to $\mathcal{C}$.) This is all that is required for Example 2.3 to be consistent with Lewis's theory.

The proof of Proposition 2.7, which clarifies the issue raised by Williamson, is essentially the same as that of completeness over standard partition models.

Proposition 2.7 (Completeness of **S5**)**. S5** is sound and complete with respect to L/D-semantics over locally absolute RA models in which $\twoheadrightarrow$ is an equivalence relation.

In general, for locally absolute RA models, the correspondence between properties of $\twoheadrightarrow$ and modal axioms is exactly as in basic modal logic.

2.9.2 Higher-Order Knowledge and Subjunctivism

The study of higher-order knowledge becomes more interesting with the subjunctivist theories, especially in connection with our primary concern of closure. According to Nozick [1981], the failures of epistemic closure implied by his tracking theory are something that "we must adjust to" (228). This would be easier if problems ended with the closure failures themselves. However, as we will see, the structural features of the subjunctivist theories that lead to these closure failures also lead to problems of higher-order knowledge.

We begin with a definition necessary for stating Fact 2.8 below.

Definition 2.12 (Outer Necessity)**.** Let us temporarily extend our language with an *outer necessity* operator $\Box$ [Lewis, 1973, §1.5] with the truth clause:

$$\mathcal{M}, w \vDash_x \Box\varphi \text{ iff } \forall v \in W_w \colon \mathcal{M}, v \vDash_x \varphi.$$

We call the language with K, B, and $\Box$ the *epistemic-doxastic-alethic* language. Define the possibility operator by $\Diamond\varphi := \neg\Box\neg\varphi$, and let $\hat{K}\varphi := \neg K\neg\varphi$.

Fact 2.8 below shows that if *sensitivity* (Def. 2.7) is necessary for knowledge, and if there is any counterfactually accessible world in which an agent believes φ but φ is false, then the agent cannot know that her belief that φ is not false—*even if she knows that φ is true.*[50] The proof appears in many places [DeRose, 1995, Kripke, 2011, Vogel, 1987, 2000, Sosa, 1996, 1999].

[50]More precisely, she cannot know that she does not have a false belief that φ [Becker, 2006]. As Becker in effect proves, $K\varphi \wedge BB\varphi \to K(B\varphi \wedge \varphi)$ is H-valid (and hence S-valid).

Fact 2.8 (Possibility and Sensitivity)**.** $\Diamond(B\varphi \wedge \neg\varphi) \rightarrow \hat{K}(B\varphi \wedge \neg\varphi)$ is H/N-valid, but not S-valid.

Since $Kp \wedge \Diamond(Bp \wedge \neg p)$ is satisfiable, $Kp \rightarrow K\neg(Bp \wedge \neg p)$ is not H/N-valid by Fact 2.8, so $Kp \rightarrow K(\neg Bp \vee p)$ is not H/N-valid. Hence Fact 2.8 is related to the failure of closure under disjunctive addition. Clearly $\Diamond\psi \rightarrow \hat{K}\psi$ is not H/N-valid for all ψ. Related to Fact 2.8, Fact 2.9 (used for Corollary 2.5) shows that limited forms of closure, including closure under disjunctive addition, hold when higher-order knowledge of $B\varphi \rightarrow \varphi$ or $\hat{K}\varphi \rightarrow \varphi$ is involved.

Fact 2.9 (Higher-Order Closure)**.**

1. $K(B\varphi \rightarrow \varphi) \rightarrow K((B\varphi \rightarrow \varphi) \vee \psi)$ is H/S-valid, but not N-valid;
2. $B\varphi \wedge K(B\varphi \rightarrow \varphi) \rightarrow K\varphi$ is H/S-valid, but not N-valid;
3. $B\varphi \wedge K(\hat{K}\varphi \rightarrow \varphi) \rightarrow K\varphi$ is H/S-valid, but not N-valid;
4. $K(\varphi \wedge \psi) \wedge K(\hat{K}\varphi \rightarrow \varphi) \rightarrow K\varphi$ is H/S-valid, but not D/N-valid;
5. $K\varphi \wedge K\psi \wedge K(\hat{K}(\varphi \vee \psi) \rightarrow (\varphi \vee \psi)) \rightarrow K(\varphi \vee \psi)$ is H/N/S-valid, but not D-valid (over total RA models).

While some consider Fact 2.8 to be a serious problem for sensitivity theories, Fact 2.10 seems even worse for subjunctivist-flavored theories in general: according to the ones we have studied, it is possible for an agent to know the classic example of an unknowable sentence, $p \wedge \neg Kp$ [Fitch, 1963]. Williamson [2000, 279] observes that $p \wedge \neg Kp$ is knowable according to the sensitivity theory. We observe that it is also knowable according to the safety theory.[51]

Fact 2.10 (Moore-Fitch Sentences)**.** $K(p \wedge \neg Kp)$ is satisfiable in RA models under D-semantics and in CB models under H/N/S-semantics.

[51] One difference between Fact 2.8 and Fact 2.10 is that the former applies to any theory for which sensitivity is a necessary condition for knowledge, whereas the latter could in principle be blocked by theories that propose other necessary conditions for knowledge in addition to sensitivity or safety. What Fact 2.10 shows is that sensitivity and safety theorists have some explaining to do about what they expect to block such a counterintuitive result.

Proof. It is immediate from Theorem 2.1 that $\neg K(p \wedge \neg Kp)$ is not D-valid.[52]

We give a simple satisfying CB model $\mathcal{M}$ for H/N/S-semantics in Fig. 2.8. Assume that $\leqslant_{w_3}$ is any appropriate preorder such that $\mathcal{M}, w_3 \vDash_{h,n,s} Kp$. It will not matter whether $w_1 \equiv_{w_1} w_2 \equiv_{w_1} w_3$ or $w_1 \equiv_{w_1} w_2 <_{w_1} w_3$.

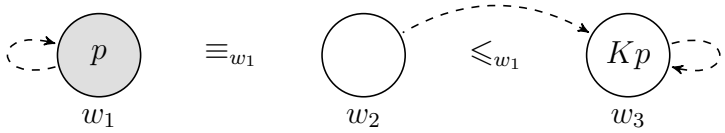

Figure 2.8: a CB model satisfying $K(p \wedge \neg Kp)$ in H/N/S-semantics (partially drawn)

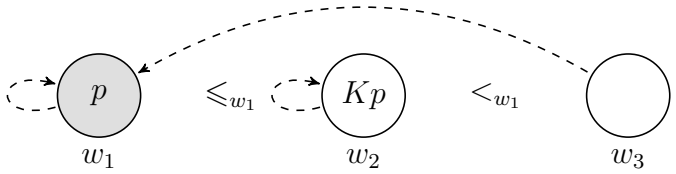

Figure 2.9: a CB model satisfying $K(p \wedge \neg Kp)$ in H/N-semantics (partially drawn)

Given $w_2 \in \mathrm{Min}_{\leqslant_{w_1}}(W)$ and $\mathcal{M}, w_2 \vDash \neg p \wedge Bp$, the safety condition for Kp fails at w_1, so $\mathcal{M}, w_1 \vDash_s p \wedge \neg Kp$. Then since $D(w_1) = \{w_1\}$ (recall Notation 2.5), $\mathcal{M}, w_1 \vDash_s B(p \wedge \neg Kp)$, so the belief condition for $K(p \wedge \neg Kp)$ holds at w_1. For $i \geq 2$, given $\mathcal{M}, w_i \vDash BKp$, we have $\mathcal{M}, w_i \nvDash B(p \wedge \neg Kp)$. It follows that for all $v \in \mathrm{Min}_{\leqslant_{w_1}}(W)$, $\mathcal{M}, v \vDash_s B(p \wedge \neg Kp) \rightarrow p \wedge \neg Kp$. Hence the safety condition for $K(p \wedge \neg Kp)$ holds at w_1, so $\mathcal{M}, w_1 \vDash_s K(p \wedge \neg Kp)$. One can check that $\mathcal{M}, w_1 \vDash_{h,n} K(p \wedge \neg Kp)$ as well. For H/N-semantics, the model $\mathcal{N}$ in Fig. 2.9, which has the same basic structure as Williamson's [2000, 279] example, also satisfies $K(p \wedge \neg Kp)$ at w_1. Assume $\leqslant_{w_2}$ is any appropriate preorder such that $\mathcal{N}, w_2 \vDash_{h,n} Kp$.[53] (Whether $w_1 \equiv_{w_1} w_2$ or $w_1 <_{w_1} w_2$ does not matter.) □

It is not difficult to tell a story with the structure of Fig. 2.8, illustrating that the

[52] Rewrite $\neg K(p \wedge \neg Kp)$ as $K(p \wedge \neg Kp) \rightarrow \bot$. T-unpacking gives $p \wedge \neg Kp \wedge K(p \wedge \neg Kp) \rightarrow \bot$ and then $p \wedge K(p \wedge \neg Kp) \rightarrow Kp$, which fails (a), (c), and (d) of Theorem 2.1.

[53] One can of course add more worlds to W_{w_2} than are shown in Fig. 2.9.

safety theory allows $K(p \wedge \neg Kp)$, just as Williamson tells a story with the structure of Fig. 2.9, illustrating that the tracking theory allows $K(p \wedge \neg Kp)$.

Fact 2.10 is related to the fact that closure under conjunction elimination is not valid. Otherwise $K(p \wedge \neg Kp)$ would be unsatisfiable; for by veridicality, $K(p \wedge \neg Kp) \rightarrow \neg Kp$ is valid, and given closure under conjunction elimination, $K(p \wedge \neg Kp) \rightarrow Kp$ would also be valid. However, Fact 2.11 shows that K does partially distribute over conjunctions of special forms in S-semantics.

Fact 2.11 (Higher-Order Closure cont.)**.**

1. $K(\varphi \wedge \neg K\varphi) \rightarrow K\neg K\varphi$ is S-valid, but not D/H/N-valid.
2. $K((B\varphi \rightarrow \varphi) \wedge (\psi \rightarrow \varphi)) \rightarrow K(B\varphi \rightarrow \varphi)$ is S-valid, but not H/N-valid.

What Facts 2.10 and 2.8 show is that in order to fully calculate the costs of closure failures, one must take into account their ramifications in the realm of higher-order knowledge. Combining Facts 2.8, 2.9, and 2.11 with results from earlier sections, we arrive at a picture of the relations between the sets of valid principles according to D-, H-, N-, and S-semantics, respectively, given by Corollary 2.5 below.[54] First we need the following definition.

Definition 2.13 (Theories and Model Classes)**.** For a class S of models, let $\mathrm{Th}^{x}_{\mathcal{L}}(\mathsf{S})$ be the set of formulas in the language $\mathcal{L}$ that are valid over S according to X-semantics. Let RAT be the class of all total RA models, RA the class of all RA models, and CB the class of all CB models.

Corollary 2.5 (Hierarchies)**.**

[54] If we require more properties of the D relation, then more principles will be valid in H/N/S-semantics—obviously for the B operator, but also for the interaction between K and B. For example, if require that D be *dense*, so $BB\varphi \rightarrow B\varphi$ is valid, then $BB\varphi \rightarrow KB\varphi$ is H/S-valid. If we also require that D be transitive, so $B\varphi \rightarrow BB\varphi$ is valid, then $B\varphi \rightarrow KB\varphi$ is H/N/S-valid. As Kripke [2011, 183] in effect observes, if $B\varphi \leftrightarrow BB\varphi$ is valid, then (for propositional φ) $\mathcal{M}, w \vDash_h K\varphi$ implies $\mathcal{M}, w \vDash_n K(\varphi \wedge B\varphi)$, so whenever $\mathcal{M}, w \vDash_h K\varphi$ but $\mathcal{M}, w \nvDash_n K\varphi$ (because adherence is not satisfied), $K(\varphi \wedge B\varphi) \rightarrow K\varphi$ fails according to N-semantics, an extreme closure failure.

1. For the flat fragment $\mathcal{L}_f$ of the epistemic language,

$$Th^n_{\mathcal{L}_f}(\mathsf{CB}) = Th^h_{\mathcal{L}_f}(\mathsf{CB}) = Th^s_{\mathcal{L}_f}(\mathsf{CB}) = Th^d_{\mathcal{L}_f}(\mathsf{RA}) \subsetneq Th^d_{\mathcal{L}_f}(\mathsf{RAT}).$$

2. For the epistemic language $\mathcal{L}_e$,

$$Th^d_{\mathcal{L}_e}(\mathsf{RA}) \subsetneq Th^n_{\mathcal{L}_e}(\mathsf{CB}) \subsetneq Th^h_{\mathcal{L}_e}(\mathsf{CB}) \subsetneq Th^s_{\mathcal{L}_e}(\mathsf{CB});$$

$$Th^d_{\mathcal{L}_e}(\mathsf{RA}) \subsetneq Th^d_{\mathcal{L}_e}(\mathsf{RAT}) \not\subseteq Th^s_{\mathcal{L}_e}(\mathsf{CB});\ Th^n_{\mathcal{L}_e}(\mathsf{CB}) \not\subseteq Th^d_{\mathcal{L}_e}(\mathsf{RAT}).$$

3. For the epistemic-doxastic language $\mathcal{L}_d$,

$$Th^n_{\mathcal{L}_d}(\mathsf{CB}) \subsetneq Th^h_{\mathcal{L}_d}(\mathsf{CB}) \subsetneq Th^s_{\mathcal{L}_d}(\mathsf{CB}).$$

4. For the epistemic-doxastic-alethic language $\mathcal{L}_a$,

$$Th^n_{\mathcal{L}_a}(\mathsf{CB}) \subsetneq Th^h_{\mathcal{L}_a}(\mathsf{CB});\ Th^n_{\mathcal{L}_a}(\mathsf{CB}) \not\subseteq Th^s_{\mathcal{L}_a}(\mathsf{CB}) \not\subseteq Th^h_{\mathcal{L}_a}(\mathsf{CB}).$$

Proof. Part 1 follows from Corollary 2.4 and Fact 2.6. Part 2 follows from Corollary 2.4, Propositions 2.5 - 2.6, and Facts 2.9.5, 2.9.4, 2.11.1, and 2.6. Part 3 follows from Propositions 2.5 - 2.6 and Facts 2.9 and 2.11. Part 4 follows from Proposition 2.5 (which clearly extends to $\mathcal{L}_a$) and Facts 2.9, 2.8, and 2.11. □

In this section we have focused on the implications of D/H/N/S-semantics for higher-order *knowledge*, especially in connection with epistemic closure. However, if we take the point of view suggested earlier (§1, §2.6, §2.8), according to which our results can be interpreted as results about desirable epistemic properties other than knowledge, then exploring higher-order phenomena in D/H/N/S-semantics is part of understanding these other properties. Along these lines, we conclude this section with an open problem for future research.

Problem 2.1 (Axiomatization). Axiomatize the theory of counterfactual belief models according to H-, N-, or S-semantics for the full epistemic, epistemic-doxastic, or

epistemic-doxastic-alethic language.

Remark 2.10 (Easy Axiomatizations)**.** If we extend the language of Definition 2.1 so that we can describe different parts of our CB models independently, e.g., by adding the belief operator B for the doxastic relation D or a counterfactual conditional $\Box\!\!\rightarrow$ for the similarity relations $\leqslant_w$, then the problem of axiomatization becomes easier. For S-semantics, which does not use the structure of any $\leqslant_w$ relation beyond $\mathrm{Min}_{\leqslant_w}(W)$, just adding B to the language makes the axiomatization problem easy. As one can prove by a standard canonical model construction, for completeness it suffices to combine the logic **KD** for B with the axiom $K\varphi \to B\varphi$ and the rule

$$\mathrm{SA}\ \frac{(B\varphi_1 \to \varphi_1) \wedge \cdots \wedge (B\varphi_n \to \varphi_n) \to (B\psi \to \psi)}{K\varphi_1 \wedge \cdots \wedge K\varphi_n \to (B\psi \to K\psi)}\,(n\geq 0).$$

For H/N-semantics, adding not only B but also a counterfactual $\Box\!\!\rightarrow$ (with the Lewisian semantics outlined in §2.5) makes the axiomatization problem easy. For example, for N-semantics we can combine **KD** for B with a complete system for counterfactuals (no interaction axioms between B and $\Box\!\!\rightarrow$ are needed), plus $K\varphi \to B\varphi$ and $K\varphi \leftrightarrow B\varphi \wedge (\neg\varphi \mathrel{\Box\!\!\rightarrow} \neg B\varphi) \wedge (B\varphi \mathrel{\Box\!\!\rightarrow} \varphi)$. The problem with obtaining easy axiomatizations by extending the language in this way is that the resulting systems give us little additional insight. The interesting properties of knowledge are hidden in the axioms that combine several operators, each with different properties. Although in a complete system for the extended language we can of course derive all principles that could appear in any sound system for a restricted language, this fact does not tell us what those principle are or which set of them is complete with respect to the restricted language. Corollary 2.4 and Facts 2.8, 2.9, and 2.11 suggest that more illuminating principles may appear as axioms if we axiomatize the S-theory of CB models in the epistemic language or the H/N-theory of CB models in the epistemic-doxastic(-alethic) language.

2.10 Theory Parameters and Closure

In this section, we return to the issue raised in §2.6.5 about the sources of closure failure. Analysis of Theorem 2.1 shows that two parameters of a modal theory of knowledge affect whether closure holds. In §2.4, we identified one: the $\forall\exists$ vs. $\exists\forall$ choice of the relevancy set. Both L- and S-semantics have an $\exists\forall$ setting of this parameter (recall Observation 2.1). However, closure holds in L-semantics but fails in S-semantics. The reason for this is the second theory parameter: the notion of ruling out. With the Lewis-style notion of ruling out in L/D-semantics, a world v is either ruled out at w or not. By contrast, with the notions of ruling out implicit in S/H/N-semantics, we cannot say independently of a proposition in question whether v is ruled out at w.

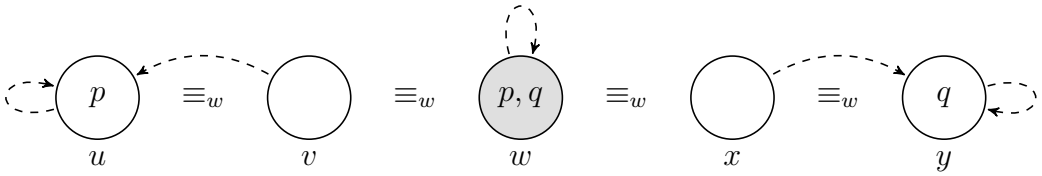

Figure 2.10: a CB countermodel for $K(p \wedge q) \rightarrow Kp \vee Kq$ in H/N/S-semantics (partially drawn)

For example, in the CB model in Fig. 2.10, v is among the closest worlds to the actual world w. We may say that v is ruled out *as an alternative for* $p \wedge q$, in the sense that while $p \wedge q$ is false at v, the agent does not believe $p \wedge q$ at v (but rather $p \wedge \neg q$). However, v is not ruled out *as an alternative for* p, for p is false at v and yet the agent believes p at v. This explains the consequence of Theorem 2.1 that $K(p \wedge q) \rightarrow Kp$ is not valid in S-semantics, because one may *safely* believe $p \wedge q$ at a world w even though one does not safely believe p at w. Note that the example also applies to sensitivity theories, for which we can again only say whether v is ruled out *as an alternative for* a given φ.

The distinction between the two notions of ruling out (RO) is again that of $\forall\exists$ vs. $\exists\forall$, as in the case of $\mathsf{RS}_{\forall\exists}$ vs. $\mathsf{RS}_{\exists\forall}$ in §2.4. Let us state the distinction in terms of possibilities that are *not* ruled out, possibilities that are *uneliminated*:

> According to an $\mathsf{RO}_{\forall\exists}$ theory, for every context $\mathcal{C}$, world w, and ($\forall$) proposition P, there is ($\exists$) a set of worlds $\mathsf{u}_{\mathcal{C}}(P,w) \subseteq \overline{P}$ *uneliminated at* w *as alternatives for* P, such that if any world in $\mathsf{u}_{\mathcal{C}}(P,w)$ is relevant (i.e., in $\mathsf{r}_{\mathcal{C}}(P,w)$), then the agent does not know P in w (relative to $\mathcal{C}$).
>
> According to an $\mathsf{RO}_{\exists\forall}$ theory, for every context $\mathcal{C}$ and world w, there is ($\exists$) a set of worlds $\mathsf{U}_{\mathcal{C}}(w)$ *uneliminated at* w, such that for every ($\forall$) proposition P, if any world in $\mathsf{U}_{\mathcal{C}}(w) \cap \overline{P}$ is relevant (i.e., in $\mathsf{r}_{\mathcal{C}}(P,w)$), then the agent does not know P in w (relative to $\mathcal{C}$).

Every $\mathsf{RO}_{\exists\forall}$ theory is a $\mathsf{RO}_{\forall\exists}$ theory (with $\mathsf{u}_{\mathcal{C}}(P,w) = \mathsf{U}_{\mathcal{C}}(w) \cap \overline{P}$), but when I refer to $\mathsf{RO}_{\forall\exists}$ theories I have in mind those that are not $\mathsf{RO}_{\exists\forall}$. As noted, L/D-semantics formalize $\mathsf{RO}_{\exists\forall}$ theories, with $\twoheadrightarrow(w)$ (Notation 2.4) in the role of $\mathsf{U}(w)$, while S/H/N-semantics formalize $\mathsf{RO}_{\forall\exists}$ theories, given the role of belief in their notions of ruling out, noted above (see §3.3.2).

Consider the parallel between $\mathsf{RS}_{\forall\exists}$ and $\mathsf{RO}_{\forall\exists}$ parameter settings: given a $\forall\exists$ setting of the RO (resp. RS) parameter, a $(\neg\varphi \wedge \neg\psi)$-world that is ruled out as an alternative for φ (resp. that must be ruled out in order to know φ) may not be ruled out as an alternative for ψ (resp. may not be such that it must be ruled out in order to know ψ), because whether the world is ruled out or not (resp. relevant or not) depends on the proposition in question, as indicated by the $\forall$ *propositions* $\exists$ *set of uneliminated (resp. relevant) worlds* quantifier order. As the example of Fig. 2.10 shows, the $\mathsf{RO}_{\forall\exists}$ setting for safety explains why closure fails in S-semantics, despite its $\mathsf{RS}_{\exists\forall}$ setting.

Table 2.2 summarizes the relationship between the two theory parameters and closure failures. Not all theories with $\mathsf{RS}_{\forall\exists}$ or $\mathsf{RO}_{\forall\exists}$ settings must have the *same* closure failures as those described in Theorem 2.1, but these settings are a good guide for when to expect closure failure, as shown by another example in §2.10.1. In Chapter 3, I will reanalyze the theories of this chapter in terms of the r and u functions to obtain a finer-grained understanding of the sources of their closure failures.

Theory	Formalization	Relevancy Set	Ruling Out	Closure Failures
RA	L-semantics	$\exists\forall$	$\exists\forall$	none
RA	D-semantics	$\forall\exists$	$\exists\forall$	Theorem 2.1
Safety	S-semantics	$\exists\forall$	$\forall\exists$	Theorem 2.1
Tracking	H/N-semantics	$\forall\exists$	$\forall\exists$	Theorem 2.1

Table 2.2: parameter settings and closure failures

2.10.1 Double-Safety

DeRose [1995] proposed an influential modification of Nozick's tracking theory, one of the perceived advantages of which was its consistency with epistemic closure. On DeRose's view, for any fixed context there is a contextually determined sphere of "epistemically relevant worlds" centered around the world w (37). To know φ at w, it must be that for any world v in that sphere, one believes φ at v iff φ is true at v (see DeRose 1995, 34 and DeRose 2004). This condition, which DeRose [2011] calls *double-safety*, is structurally equivalent to the combination of safety and adherence.

We can represent DeRose's double-safety condition in CB models by taking $\leqslant_w$ to be a relevance ordering, as in RA models, and by taking the epistemically relevant worlds around w to be those in $\mathrm{Min}_{\leqslant_w}(W)$.[55] DeRose's full contextualist view involves both a set of epistemically relevant worlds and a similarity ordering of worlds for counterfactuals. When context changes, worlds are added to the set of worlds that are epistemically relevant at w, based on their distance from w according to the similarity ordering. However, the point I wish to make here is that closure fails for double-safety in a *fixed context*, which we can see without representing the relevance and similarity orderings separately. To model DeRose's contextualism formally, we can use techniques similar to those used to model Lewis's contextualism in §2.11, applied to combined RA-CB models that include both relevance and similarity orderings.

[55] As with safety, the ordering of worlds beyond the (most) relevant worlds will not make a difference in the truth definition for $K\varphi$ in a fixed context, so we could instead define a function e that assigns to each world w a set $e(w)$ of epistemically relevant worlds with $w \in e(w)$. The result with $e(w)$ in place of $\mathrm{Min}_{\leqslant_w}(W)$ would then be equivalent to Definition 2.14 below.

I call the semantics with the double-safety condition R-semantics for De**R**ose.

Definition 2.14. Given a CB model $\mathcal{M} = \langle W, D, \leqslant, V\rangle$ and $w \in W$, we define $\mathcal{M}, w \vDash_r \varphi$ as follows (with other clauses as in Definition 2.7):

$$\mathcal{M}, w \vDash_r K\varphi \quad \text{iff} \quad \mathcal{M}, w \vDash_r B\varphi \text{ and}$$
$$\text{(double safety) } \forall v \in \mathrm{Min}_{\leqslant_w}\big(W\big)\colon \mathcal{M}, v \vDash_r B\varphi \leftrightarrow \varphi.$$

Like safety, double-safety combines an $\mathsf{RS}_{\exists\forall}$ choice of the relevancy set with a $\mathsf{RO}_{\forall\exists}$ notion of ruling out.[56] According to the explanation of closure failure above, we should expect that closure fails in R-semantics. This expectation is in conflict with the assumption in some of the literature that closure holds for double-safety, which is perhaps due to the implicit assumption that an $\mathsf{RS}_{\exists\forall}$ choice of the relevancy set guarantees closure.[57] However, given the $\mathsf{RO}_{\forall\exists}$ notion of ruling out, closure does fail in R-semantics. Just as the model in Figure 2.3 shows that one's belief that $p \wedge q$ can be safe at w even though one's belief that q is not, the model also shows that one's belief that $p \wedge q$ can be *double-safe* at w even though one's belief that q is not.

All of the other closure principles we have shown to fail in D/H/N/S-semantics (Facts 2.2 and 2.3 and Exercise 2.1) also fail in R-semantics, as one can easily check. Indeed, the double-safety theory suffers from the same major closure failures as Nozick's theory. Since one of the perceived advantages of double-safety over tracking was an ability to avoid these closure failures, this is a negative result. Moreover, there are other problems with double-safety and safety to be discussed in §4.1.

2.11 The Dynamics of Context

In §2.4, I remarked that contextualists should think of an RA model $\mathcal{M}$ as associated with a fixed context of knowledge attribution, so a change in context corresponds to

[56] R-semantics is to N-semantics what S-semantics is to H-semantics: just as safety is what one obtains by restricting sensitivity to a fixed set of worlds (recall §2.7), double-safety is what one obtain by restricting full tracking (sensitivity plus adherence) to a fixed set of worlds.

[57] Doubts about whether closure holds for double-safety have been raised by Cohen [1999, 72f] and Vogel [2007, 87]. However, it seems that the key reason for the failure of closure for double-safety, explained in the text, was not identified.

a change in models from $\mathcal{M}$ to $\mathcal{M}'$. In this section, I will make this idea precise.[58]

In the framework of Lewis [1979], the family $\preceq$ of relevance orderings in an RA model may be thought of as a component of the *conversational score*. Changes in this component of the conversational score, an aspect of what Lewis calls the *kinematics of score*, correspond to transformations of RA models. We begin with an RA model $\mathcal{M}$ representing what an agent counts as knowing relative to an initial conversational context. If some change in the conversation makes the issue of φ relevant, then corresponding to this change the model transforms from $\mathcal{M}$ to $\mathcal{M}^{\uparrow\varphi}$. In the new model, what the agent counts as knowing may be different.

For variety, we will define two types of operations on models, $\uparrow \varphi$ and $\Uparrow \varphi$. Roughly speaking, $\uparrow \varphi$ changes the model so that the *most relevant φ-worlds* in $\mathcal{M}$ become among the *most relevant worlds overall* in $\mathcal{M}^{\uparrow\varphi}$. By contrast, $\Uparrow \varphi$ changes the model so that any worlds *at least as relevant as* the most relevant φ-worlds in $\mathcal{M}$ become among the most relevant worlds overall in $\mathcal{M}^{\Uparrow\varphi}$. The following definition makes these descriptions more precise. For convenience, in this section we assume that each preorder $\preceq_w$ is total on its field W_w, but all of the definitions and results can be modified to apply to the non-total case.

Definition 2.15 (RA Context Change)**.** Given an RA model $\mathcal{M} = \langle W, \rightarrow, \preceq, V\rangle$, define the models $\mathcal{M}^{\uparrow\varphi} = \left\langle W, \rightarrow, \preceq^{\uparrow\varphi}, V\right\rangle$ and $\mathcal{M}^{\Uparrow\varphi} = \left\langle W, \rightarrow, \preceq^{\Uparrow\varphi}, V\right\rangle$ such that for all $w, u, v \in W$:

1. if $u \in \mathrm{Min}_{\preceq_w}\left(\llbracket\varphi\rrbracket^{\mathcal{M}}\right) \cup \mathrm{Min}_{\preceq_w}(W)$, then $u \preceq^{\uparrow\varphi}_w v$;

2. if $u, v \notin \mathrm{Min}_{\preceq_w}\left(\llbracket\varphi\rrbracket^{\mathcal{M}}\right) \cup \mathrm{Min}_{\preceq_w}(W)$, then $u \preceq^{\uparrow\varphi}_w v$ iff $u \preceq_w v$;

and

3. if $\exists x \in \mathrm{Min}_{\preceq_w}\left(\llbracket\varphi\rrbracket^{\mathcal{M}}\right)$ such that $u \preceq_w x$, then $u \preceq^{\Uparrow\varphi}_w v$;

4. if $\forall x \in \mathrm{Min}_{\preceq_w}\left(\llbracket\varphi\rrbracket^{\mathcal{M}}\right)$, $u \npreceq_w x$ and $v \npreceq_w x$, then $u \preceq^{\Uparrow\Phi}_w v$ iff $u \preceq_w v$.

[58]The material from this section and §2.F is drawn from my "Epistemic Logic, Relevant Alternatives, and the Dynamics of Context" [Holliday, 2012].

In other words, for $\uparrow \varphi$, the *most relevant φ-worlds* according to $\preceq_w$ become among the *most relevant worlds* according to $\preceq_w^{\uparrow\varphi}$; the most relevant worlds according to $\preceq_w$ remain among the most relevant worlds according to $\preceq_w^{\uparrow\varphi}$; and for all other worlds, $\preceq_w^{\uparrow\varphi}$ agrees with $\preceq_w$. For $\Uparrow \varphi$, all worlds *at least as relevant as* the most relevant φ-worlds according to $\preceq_w$ become among the most relevant worlds according to $\preceq_w^{\Uparrow\varphi}$; and for all other worlds, $\preceq_w^{\Uparrow\varphi}$ agrees with $\preceq_w$.

Which of these operations is most appropriate for modeling a given context change is an interesting question, which I leave aside here. Other operations could be defined as well, but these will suffice as examples of the general method. Fig. 2.11 shows the application of either $\uparrow x$ or $\Uparrow x$ (denoted $+x$) to the model $\mathcal{M}$ for Example 1.1, the result of which is the same for both. Fig. 2.12 shows $\uparrow x$ and $\Uparrow x$ applied to a different initial model, $\mathcal{N}$, in which case the results are different.

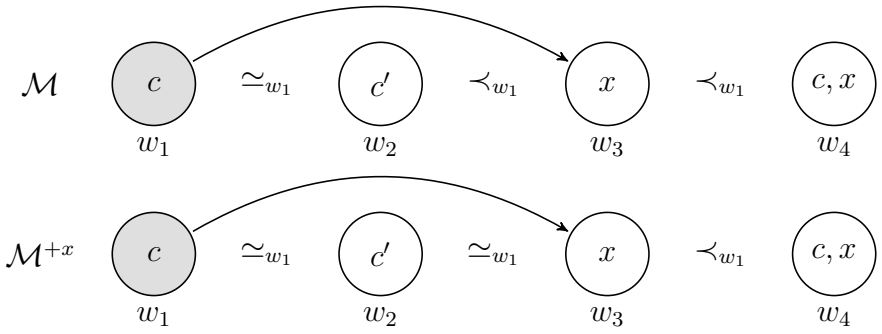

Figure 2.11: result of context change by raising x in Example 1.1

To describe the effect of these context change operations using our formal language, we extend the language of Definition 2.1 with *dynamic* context change operators of the form $[+\varphi]$ for $+ \in \{\uparrow, \Uparrow\}$, in the style of dynamic epistemic logic [van Ditmarsch et al., 2008, Benthem, 2011]. One can read $[+\varphi]\psi$ as "after φ becomes relevant, ψ is the case" or "after φ is raised, ψ is the case" or "after context change by φ, ψ is the case," etc.

Definition 2.16 (Contextualist Epistemic Language)**.** Let $\mathsf{At} = \{p, q, r, \dots\}$ be a set of atomic sentences. The *contextualist epistemic language* is generated as follows,

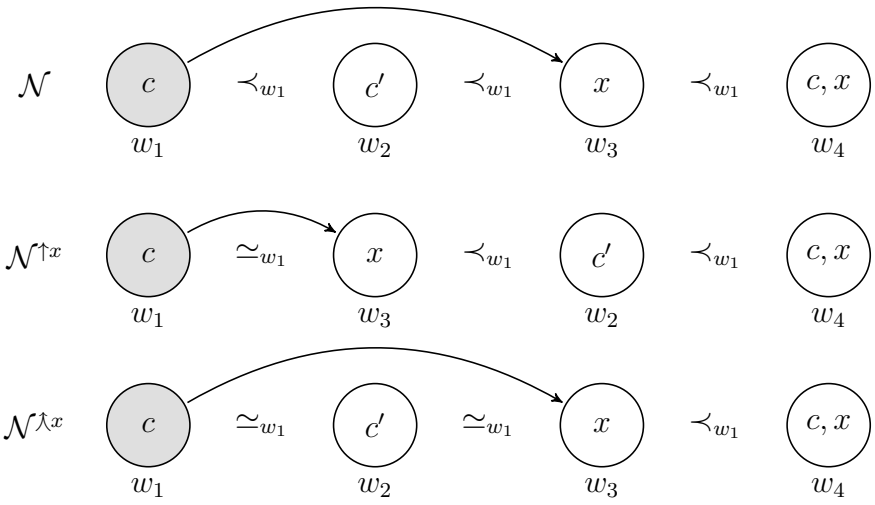

Figure 2.12: different results of context change by $\uparrow x$ and $\Uparrow x$

where $p \in \mathsf{At}$:

$$\begin{aligned} \varphi &::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K\varphi \mid [\pi]\varphi \\ \pi &::= \uparrow \varphi \mid \Uparrow \varphi. \end{aligned}$$

We give the truth clauses for the operators $[\uparrow \varphi]$ and $[\Uparrow \varphi]$ with the help of Definition 2.15, using $+$ to stand for either $\uparrow$ or $\Uparrow$ in definitions applicable to both.

Definition 2.17 (Truth)**.** The truth clause for the context change operator is:

$$\mathcal{M}, w \vDash [+\varphi]\psi \text{ iff } \mathcal{M}^{+\varphi}, w \vDash \psi.$$

In other words, "after context change by φ, ψ is the case" is true at w in the initial model $\mathcal{M}$ if and only if ψ is true at w in the new model $\mathcal{M}^{+\varphi}$.

Having set up this contextualist machinery, there are a number of directions to explore. For the purposes of my argument, the most important is a comparison between (non-contextualist) D-semantics and contextualist L-semantics. Appendix §2.F contains a technical excursion in search of *reduction axioms* for context change.

2.11.1 D-Semantics vs. Contextualist L-Semantics

The following fact matches Lewis's [1996] view on closure and context discussed in §2.1. (It may be helpful to reread the relevant part of §2.1 before proceeding here.)

Fact 2.12 (Known Implication Cont.)**.** According to D-semantics, closure under known implication can fail. According to L-semantics, closure under known implication always holds for a fixed context, but may fail across context changes:

1. $\not\models_d K\varphi \wedge K(\varphi \to \psi) \to K\psi$
2. $\models_l K\varphi \wedge K(\varphi \to \psi) \to K\psi$
3. $\not\models_l K\varphi \wedge K(\varphi \to \psi) \to [+\neg\psi]K\psi$
4. $\not\models_l K\varphi \to [+\neg\psi](K(\varphi \to \psi) \to K\psi)$

Proof. We have already noted part 1 and 2 in §2.4. For 3, its instance

$$Kc \wedge K(c \to \neg x) \to [+x]K\neg x \tag{2.56}$$

is false at $\mathcal{M}, w_1$ in Fig. 2.11. As we saw in §2.4, the antecedent is true at $\mathcal{M}, w_1$. To determine whether $\mathcal{M}, w_1 \models_l [+x]K\neg x$, by Definition 2.17 we must check whether $\mathcal{M}^{+x}, w_1 \models_l K\neg x$. Since in $\mathcal{M}^{+x}$ there is a most relevant (at w_1) world, w_3, which satisfies x and is not ruled out at w_1, we have $\mathcal{M}^{+x}, w_1 \not\models_l K\neg x$. Therefore, $\mathcal{M}, w_1 \not\models_l [+x]K\neg x$, so (2.56) is false at $\mathcal{M}, w_1$. It is also easy to check that $\mathcal{M}^{+x}, w_1 \models K(c \to \neg x)$, so the corresponding instance of 4 is false at $\mathcal{M}, w_1$. □

We will use the next fact to generalize Fact 2.12 to all kinds of closure failure (Fact 2.14), not only failures of closure under known implication.

Fact 2.13 (Relation of D- to Contextualist L-semantics)**.** Given an RA model $\mathcal{M} = \langle W, \rightarrowtriangle, \preceq, V\rangle$ with $w \in W$, for any *propositional* formula φ,

$$\mathcal{M}, w \models_d K\varphi \text{ iff } \mathcal{M}, w \models_l [+\neg\varphi]K\varphi.$$

Proof. For the case where $+$ is $\uparrow$, by Definition 2.15,

$$\mathrm{Min}_{\preceq_w^{\uparrow\neg\varphi}}(W) = \mathrm{Min}_{\preceq_w}(W) \cup \mathrm{Min}_{\preceq_w}(\overline{[\![\varphi]\!]}^{\mathcal{M}}), \tag{2.57}$$

so

$$\mathrm{Min}_{\preceq_w^{\uparrow\neg\varphi}}(W) \cap \overline{[\![\varphi]\!]}^{\mathcal{M}^{\uparrow\neg\varphi}} = (\mathrm{Min}_{\preceq_w}(W) \cup \mathrm{Min}_{\preceq_w}(\overline{[\![\varphi]\!]}^{\mathcal{M}})) \cap \overline{[\![\varphi]\!]}^{\mathcal{M}^{\uparrow\neg\varphi}}. \tag{2.58}$$

Since φ is propositional, by an easy induction we have

$$\overline{[\![\varphi]\!]}^{\mathcal{M}^{+\neg\varphi}} = \overline{[\![\varphi]\!]}^{\mathcal{M}}, \tag{2.59}$$

so from (2.58) we have

$$\begin{aligned} \mathrm{Min}_{\preceq_w^{\uparrow\neg\varphi}}(W) \cap \overline{[\![\varphi]\!]}^{\mathcal{M}^{\uparrow\neg\varphi}} &= (\mathrm{Min}_{\preceq_w}(W) \cup \mathrm{Min}_{\preceq_w}(\overline{[\![\varphi]\!]}^{\mathcal{M}})) \cap \overline{[\![\varphi]\!]}^{\mathcal{M}} \\ &= \mathrm{Min}_{\preceq_w}(\overline{[\![\varphi]\!]}^{\mathcal{M}}). \end{aligned} \tag{2.60}$$

It follows from (2.60) that

$$\mathrm{Min}_{\preceq_w}(\overline{[\![\varphi]\!]}^{\mathcal{M}}) \cap {\rightarrow\!\!\!\!\triangleright}(w) = \emptyset \tag{2.61}$$

is equivalent to

$$\mathrm{Min}_{\preceq_w^{\uparrow\neg\varphi}}(W) \cap \overline{[\![\varphi]\!]}^{\mathcal{M}^{\uparrow\neg\varphi}} \cap {\rightarrow\!\!\!\!\triangleright}(w) = \emptyset, \tag{2.62}$$

which by Definition 2.5 means that $\mathcal{M}, w \vDash_d K\varphi$ is equivalent to $\mathcal{M}^{\uparrow\neg\varphi}, w \vDash_l K\varphi$, which by Definition 2.17 is equivalent to $\mathcal{M}, w \vDash_l [\uparrow \neg\varphi]K\varphi$.

The proof for the case where $+$ is $\Uparrow$ is similar. □

Using Fact 2.13, we can now state a generalization of Fact 2.12 as follows.

Fact 2.14 (Inter-context Closure Failure)**.** Let $\varphi_1, \ldots, \varphi_n$ and ψ be propositional formulas. Given an RA model $\mathcal{M} = \langle W, {\rightarrow\!\!\!\!\triangleright}, \preceq, V\rangle$ with $w \in W$, if

$$\mathcal{M}, w \nvDash_d K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi$$

then

$$\mathcal{M}, w \not\models_l K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow [+\neg\psi]K\psi.$$

Proof. Assume the first line. Since for any formula φ, $\mathcal{M}, w \models_d K\varphi$ implies $\mathcal{M}, w \models_l K\varphi$, we have $\mathcal{M}, w \models_l K\varphi_1 \wedge \cdots \wedge K\varphi_n$. Since $\mathcal{M}, w \not\models_d K\psi$, we have $\mathcal{M}, w \not\models_l [+\neg\psi]K\psi$ by Fact 2.13, which gives the second line. □

Remark 2.11. Most contextualists deny that closure fails in any of the ways allowed by D-semantics, as described by Theorem 2.1. But Fact 2.14 shows that for *every way* in which closure fails for D-semantics, there is a corresponding *inter-context* "closure failure" for L-semantics when the context changes with the negation of the consequent of the closure principle becoming relevant. According to some standard contextualist views, asserting that the agent knows the consequent has just this effect on the context. For example, according to DeRose [1995, 37], "When it's asserted that S knows (or doesn't know) that P, then, if necessary, enlarge the sphere of epistemically relevant worlds so that it at least includes the closest worlds in which P is false." According to Lewis [1996, 559], "No matter how far-fetched a certain possibility may be, no matter how properly we might have ignored it in some other context, if in this context we are not in fact ignoring it but attending to it, then for us now it is a relevant alternative." I will return to these ideas in §4.1.2.

2.12 Conclusion

In this chapter, our model-theoretic approach helped to illuminate the structural features of RA and subjunctivist theories that lead to closure failure, as well as the precise extent of their closure failures in Theorem 2.1.

When understood as theories of *knowledge*, the basic subjunctivist-flavored theories formalized by D/H/N/S-semantics have a bad balance of closure properties, invalidating very plausible closure principles (recall §2.6) while validating questionable ones (recall §2.8). The theories formalized by C- and L-semantics also have their problems. On the one hand, the idea that knowledge requires ruling out *all* possibilities of error, reflected in C-semantics, makes knowing too hard, giving us the problem

of *skepticism* (recall §2.1 - 2.4). On the other hand, the idea that knowledge of contingent empirical truths can be acquired without ruling out *any* possibilities of error, reflected in L-semantics (and S-semantics), seems to make knowing too easy, giving us the problem of *vacuous knowledge* (recall §2.4 - 2.5). An attraction of D/H/N-semantics is that they avoid these problems. But they do so at a high cost when it comes to closure.

In Chapter 3, I will generalize the RA and CB frameworks to obtain an even finer-grained analysis of their properties. Later, in Chapter 5, I will propose a new picture of knowledge that avoids the problems of skepticism and vacuous knowledge, without the high-cost closure failures of the subjunctivist-flavored theories. As we shall see, the model-theoretic epistemic-logical approach followed here can help us not only to better understand epistemological problems, but also to discover possible solutions.

The results of this chapter motivate some methodological reflections on our approach. In epistemology, a key method of theory assessment involves considering the verdicts issued by different theories about which knowledge claims are true in a particular scenario. This is akin to considering the verdicts issued by different semantics about which epistemic formulas are true in a particular model. All of the semantics we studied can issue different verdicts for the same model. Moreover, theorists who favor different theories/semantics may represent a scenario with different models in the first place. Despite these differences, there are systematic relations between the RA, tracking, and safety perspectives represented by our semantics. In several cases, we have seen that any model viewed from one perspective can be transformed into a model that has an equivalent epistemic description from a different perspective (Propositions 2.4 - 2.6). As we have also seen, when we rise to the level of truth in *all models*, of validity, differences may wash away, revealing unity on a higher level. Theorem 2.1 provided such a view, showing that four different epistemological pictures validate essentially the same epistemic closure principles. Against this background of similarity, subtle differences within the RA/subjunctivist family appear more clearly. The picture offered by total relevant alternatives models lead to a *logic of ranked relevant alternatives*, interestingly different from the others (Corollary 2.4). In the realm of higher-order knowledge, there emerged hierarchies in the strength of different

theories (Corollary 2.5).

For some philosophers, a source of hesitation about epistemic logic is the degree of idealization. In basic systems of epistemic logic, agents know all the logical consequences of what they know, raising the "problem of logical omniscience" noted in §1. However, in our setting, logical omniscience is a feature, not a bug. Although in our formalizations of the RA and subjunctivist theories, agents do not know all the logical consequences of what they know, due to failures of epistemic closure, they are still logically omniscient in another sense. For as "ideally astute logicians" (recall §2.1), they know all logically valid principles, and they believe all the logical consequences of what they believe. These assumptions allow us to distinguish failures of epistemic closure that are due to fact that finite agents do not always "put two and two together" from failures of epistemic closure that are due to the special conditions on knowledge posited by the RA and subjunctivist theories.[59] This shows the positive role that idealization can play in epistemology, as it does in science.

2.A Comparison with Basic Epistemic Logic

In this Appendix, we compare D- and L-semantics with the standard semantics for epistemic logic, which is formally the same as C-semantics. Some important distinctions for the conceptual foundations of epistemic logic become clear in the comparison.

We will make two related comparisons: we will compare the RA framework of §2.4 with the framework of basic epistemic logic; and we will compare the notion of an *uneliminated* possibility with that of an epistemically *accessible* possibility and that of (what I call) an epistemically *live* possibility.

Models for basic epistemic logic are tuples $\mathcal{M} = \langle W, E, V\rangle$, where W and V are as in Definition 2.2, and E is a binary relation on W, required to be at least reflexive. Intuitively, we take wEv to mean that v is *epistemically accessible* from w, in the sense that everything the agent *knows* in w is *true* in v (see, e.g., Lewis 1986, §1.4,

[59]Recall note 23. Williamson [2010, 256] makes a similar point, namely that it can be useful to assume logical omniscience in order to discern the specific epistemic effects of limited powers of perceptual discrimination, as opposed to limited logical powers.

Williamson 2000, §8.2, Williamson 2009, 21), which is clearly a reflexive relation. The truth clause for $K\varphi$ is then given by:

$$\mathcal{M}, w \vDash K\varphi \text{ iff } \forall v \in W\text{: if } wEv\text{, then } \mathcal{M}, v \vDash \varphi. \tag{2.63}$$

One might sense some circularity or triviality in defining the truth of $K\varphi$ in terms of the relation E, given that we take wEv to mean that everything the agent *knows* in w is true in v. Technically, there is no circularity, because E is a primitive in the model, not defined in terms of anything else. Conceptually, one must be clear about the role of a basic epistemic model when paired with (2.63): its role is to represent the content of one's knowledge, *what one knows*, not to analyze *what knowledge is* in terms of something else. Furthermore, (2.63) is not trivial because it is not neutral with respect to all theories of knowledge. By basic results (see Theorem 3.3(2) of Chellas 1980), all closure principles are valid according to (2.63), so (2.63) excludes theories that allow closure failure. The left-to-right direction of (2.63) is neutral, for it is immediate from our notion of epistemic accessibility that if wEv and φ is false in v, then the agent does not know φ in w. However, the right-to-left direction of (2.63) is not immediate from the notion of epistemic accessibility.

Let us now compare basic epistemic semantics with the RA semantics of §2.4. To do so, we will define epistemic accessibility relations within RA models.

Definition 2.18 (Accessible)**.** While the E relation in basic epistemic models is a primitive, given an RA model $\mathcal{M} = \langle W, \rightarrow\!\!\triangleright, \preceq, V\rangle$, we define a derived epistemic accessibility relation E_x on W as follows:

$$wE_xv \text{ iff } \forall\varphi\text{: if } \mathcal{M}, w \vDash_x K\varphi\text{, then } \mathcal{M}, v \vDash_x \varphi.$$

There are two important observations to be made about this definition. First, as shown below, the E_d and E_l relations may be *different* than the $\rightarrow\!\!\triangleright$ relation; hence the set of epistemically accessible worlds and the set of *uneliminated* worlds may be distinct. Second, clause (2.63) applied to E_d and the D-semantics clause of Definition 2.5 may assign different truth values to the same formula $K\varphi$.

For example, in the model $\mathcal{M}$ in Figure 2.13, we have $w_1 \rightarrowtriangle w_3$ but *not* $w_1 E_d w_3$ or $w_1 E_l w_3$, since $\mathcal{M}, w_1 \vDash_{d,l} Kc$ but $\mathcal{M}, w_3 \nvDash_{d,l} c$. Hence for the RA theory, there can be *uneliminated possibilities that are not epistemically accessible*. One can also check that for all $v \in W$ such that $w_1 E_d v$, $\mathcal{M}, v \vDash \neg x$. Hence $\mathcal{M}, w_1 \vDash K\neg x$ according to (2.63). Yet as we have seen, $\mathcal{M}, w_1 \nvDash_d K\neg x$.

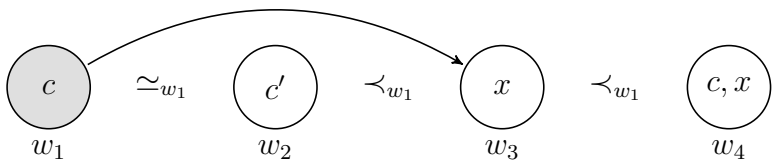

Figure 2.13: RA model for Example 1.1 (partially drawn, reflexive loops omitted)

Before considering the conceptual significance of these observations, let us introduce one more distinction. I will say that v is *epistemically live* for the agent in w iff the agent does not know in w that possibility v does not obtain, where this is understood as follows. Let us assume that we are dealing with RA models in which each world u is uniquely definable by a formula φ_u of the epistemic language.

Definition 2.19 (Live)**.** Given an RA model $\mathcal{M} = \langle W, \rightarrowtriangle, \preceq, V \rangle$, we can define an *epistemic liveness* relation L_x on W as follows:

$$w L_x v \text{ iff } \mathcal{M}, w \nvDash_x K\neg\varphi_v.$$

It may seem that the *live* worlds should be exactly the *accessible* worlds or perhaps the *uneliminated* worlds. In fact, the relations are more interesting, as illustrated by the following proposition. I leave the proof to the reader.

Proposition 2.8 (Comparing Accessible, Live, and Uneliminated)**.**

1. For all RA models, $E_d \subseteq L_d = \rightarrowtriangle$, but there are RA models with $E_d \subsetneq L_d$.
2. For all RA models, $E_l = L_l \subseteq \rightarrowtriangle$, but there are RA models with $L_l \subsetneq \rightarrowtriangle$.

Proposition 2.8 shows that if we go beyond basic epistemic logic as we have, then we must keep apart three different notions that are easily conflated.

For part 1, that $L_d = \twoheadrightarrow$ shows that a D-semantical theorist can take *eliminating* a possibility to be (extensionally) equivalent to knowing that the possibility does not obtain. However, allowing $E_d \subsetneq L_d$ means allowing that in w an agent can know something that is false in v (so v is not accessible) without knowing that v does not obtain (so v is live/uneliminated). This is another way of seeing closure failure: at the pointed model $\mathcal{M}, w_1$ in Figure 2.13 for Example 1.1, according to D-semantics student `A` knows c, which is false at w_3, but she does not know that w_3 does not obtain, because she does not know $\neg(x \wedge \neg c)$, which uniquely defines w_3.

For part 2, that $E_l = L_l$ is another expression of the fact that closure holds in L-semantics. However, since the semantics allows $L_l \subsetneq \twoheadrightarrow$, an L-semantical theorist cannot take eliminating a possibility to be equivalent to knowing that the possibility does not obtain. Pryor [2001, 99] also observes that such an equivalence is not available to RA theorists who wish to maintain closure. They require an independent notion of eliminating or ruling out a possibility, such as Lewis's notion involving perceptual experience and memory or the alternative notion that Pryor suggests.

2.B Closest vs. Close Enough

In Definition 2.7, I stated the sensitivity, adherence, and safety conditions using the $\mathrm{Min}_{\leqslant_w}$ operator, which when applied to a set S of worlds gives the set of "closest" worlds to w out of those in S. This appears to conflict with the views of Heller [1989, 1999a], who argues for a "close *enough* worlds" analysis rather than a "clos*est* worlds" analysis for sensitivity, and of Pritchard [2005, 72], who argues for considering *nearby* rather than only *nearest* worlds for safety and sensitivity. However, the conflict is merely apparent. For if one judges that the clos*est* worlds in a set S, according to $\leqslant_w$, do not include all of the worlds in S that are close *enough*, then we can relax $\leqslant_w$ to a coarser preorder $\leqslant'_w$, of which $\leqslant_w$ is a refinement, so that the closest worlds in S according to $\leqslant'_w$ are exactly those worlds in S previously judged to be closest or close enough.

To be precise, given a set $\mathit{CloseEnough}(w) \subseteq W_w$ such that $\mathrm{Min}_{\leqslant_w}(W) \subseteq \mathit{CloseEnough}(w)$ and such that if $y \in \mathit{CloseEnough}(w)$ and $x \leqslant_w y$, then $x \in$

$CloseEnough(w)$, define $\leqslant'_w$ as follows: $v \leqslant'_w u$ iff either $v \leqslant_w u$ or [$u \leqslant_w v$ and $v \in CloseEnough(w)$]. Then $\mathrm{Min}_{\leqslant'_w}(S) = \mathrm{Min}_{\leqslant_w}(S) \cup (CloseEnough(w) \cap S)$, so the close enough S-worlds are included, as desired. For the coarser preorder $\leqslant'_w$, $\mathrm{Min}_{\leqslant'_w}(W) = CloseEnough(w)$ would be the set of worlds close enough/nearby to w. Here we assume, following Heller [1999a, 201f], that whether a world counts as *close enough/nearby* may be context dependent, but for a fixed context, whether a world is close enough/nearby is not relative to the φ for which we are assessing $K\varphi$ (cf. Cross 2008 on counterfactual conditionals and antecedent-relative comparative world similarity); as discussed in §2.1, the fact that (for a given world) there is a single, fixed ordering on the set of worlds is what Heller [1999a] uses to reply to Stine's [1976] equivocation charge against Dretske. Finally, note that while the coarser preorder $\leqslant'_w$ may not be the appropriate relation for assessing counterfactuals, according to the Heller/Pritchard view, it would be appropriate for assessing knowledge.

2.C Necessary Conditions and Closure Failures

Let us return to the issue raised in Remark 2.4 about the relation between closure failures for a necessary condition of knowledge and closure failures for knowledge itself. Suppose that C is a necessary but insufficient condition for knowledge, and let $C\varphi$ mean that the agent satisfies C with respect to φ. Hence $K\varphi \rightarrow C\varphi$ should be valid. Further suppose that

$$C\varphi_1 \wedge \cdots \wedge C\varphi_n \rightarrow C\psi \tag{2.64}$$

is not valid. As Vogel [1987] and Warfield [2004] point out, it does not follow that

$$K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi \tag{2.65}$$

is not valid. For in the counterexample to (2.64), $K\varphi_1 \wedge \cdots \wedge K\varphi_n$ may not hold, since C is not sufficient for K.

Let C' be any other condition such that C and C' are jointly sufficient for K, so $C\varphi \wedge C'\varphi \rightarrow K\varphi$ is valid. If (2.65) is valid, then $C'\varphi_1 \wedge \cdots \wedge C'\varphi_n$ does not hold in

the counterexample to (2.64). Moreover, it must be that while (2.64) is not valid,

$$(C\varphi_1 \wedge \cdots \wedge C\varphi_n \wedge C'\varphi_1 \wedge \cdots \wedge C'\varphi_n) \rightarrow C\psi \tag{2.66}$$

is valid. For if there is a counterexample to (2.66), then there is a counterexample to (2.65), since C and C' are jointly sufficient and C is necessary for K.

The problem is that proposed conditions for K are typically independent in such a way that assuming one also satisfies C' with respect to $\varphi_1, \ldots, \varphi_n$ will not guarantee that one satisfies a distinct, non-redundant condition C with respect to ψ, if satisfying C with respect to $\varphi_1, \ldots, \varphi_n$ is not already sufficient. For example, if ruling out the relevant alternatives to $\varphi_1, \ldots, \varphi_n$ is not sufficient for ruling out the relevant alternatives to ψ, then what other condition is such that also satisfying it with respect to $\varphi_1, \ldots, \varphi_n$ will guarantee that one has ruled out the relevant alternatives to ψ? The same question arises for subjunctivist conditions. It is up to subjunctivists to say what they expect to block closure failures for knowledge, given closure failures for their necessary subjunctivist conditions on knowledge.

One way to do so is to build in the satisfaction of closure itself as another necessary condition. For example, Luper-Foy [1984, 45n38] gives the "trivial example" of *contracking* φ, which is the condition (C') of satisfying the sensitivity condition (C) for all logical consequences of φ. However, this idea for building in closure misses the fact that multi-premise closure principles fail for contracking. For example, one can contrack p and contract q, while being *insensitive* with respect to $(p \wedge q) \vee r$ and therefore failing to contrack $p \wedge q$.

Contracking must be distinguished from another idea for combining tracking with closure. Roush [2005, Ch. 2, §1] proposes a disjunctive account according to which (to a first approximation) an agent knows ψ iff either the agent "Nozick-knows" ψ, i.e., satisfies Nozick's belief, sensitivity, and adherence conditions for ψ, or there are some $\varphi_1, \ldots, \varphi_n$, none of which is equivalent to ψ, such that the agent knows $\varphi_1, \ldots, \varphi_n$ and knows that $\varphi_1 \wedge \cdots \wedge \varphi_n$ implies ψ. Importantly, according to this *recursive tracking view of knowledge*, the tracking conditions (for which closure fails) are not necessary conditions for knowledge.

2.D Bases and Methods

In this section, we consider the closure properties of knowledge according to well-known versions of the tracking and safety theories that take into account an agent's *method* of coming to believe or *basis* for believing a proposition.[60] Suppose an agent comes to believe φ by some method m. Call her belief that φ *sensitive** iff in the closest $\neg\varphi$-worlds, she does not believe φ *by method* m (though she may believe φ by some $m' \neq m$).[61] (For simplicity, here I ignore adherence, although the treatment below easily extends to adherence.) Similarly, suppose an agent comes to believe φ on some basis b. Call her belief that φ *safe** iff in all close worlds, if φ is false, then she does not believe φ *on basis* b (though she may believe φ on some $b' \neq b$).

To study closure for the sensitivity* and safety* theories, I will define a language inspired by that of *justification logic* [Artemov, 2008, Fitting, 2009]. The definitions will refer to 'bases', but I intend this to cover methods as well.

Definition 2.20 (Basis Language)**.** Let $\mathsf{At} = \{p, q, r, \dots\}$ be a set of atomic sentences and $\mathsf{Ba} = \{a, b, c, \dots\}$ a set of atomic bases (or methods). The epistemic-doxastic *basis language* is defined inductively by

$$\begin{aligned} b &::= a \mid b \cdot b \\ \varphi &::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid B(\varphi, b) \mid B\varphi \mid K(\varphi, b) \mid K\varphi, \end{aligned}$$

where $a \in \mathsf{Ba}$ and $p \in \mathsf{At}$.

Let us take $B(\varphi, b)$ to mean that the agent believes φ on basis b and $K(\varphi, b)$ to mean that the agent knows φ on basis b, which is to say that she believes φ on basis

[60] The motivation for taking account of methods of belief formation in the tracking theory was not to restore closure, but to block counterexamples to the theory [Nozick, 1981, 179ff]. By contast, one of the motivations for taking into account bases of beliefs in the safety theory was to restore closure (see, e.g., Sosa 1999, 149, Williamson 2009, 20).

[61] Here we follow Luper-Foy's [1984] statement of the sensitivity condition with methods, which differs from that of Nozick [1981]. According to Nozick's version, sensitivity* requires that in the closest worlds where both φ is false *and* the agent uses method m to determine *whether or not* φ, the agent does not come to believe φ by method m. This statement assumes what Luper-Foy calls "two-sided methods ... capable of recommending the belief that not-p as well as the belief that p," whereas Luper-Foy's revised version handles methods that are "one-sided, i.e., not capable of recommending the belief that not-p" (28).

b and her belief constitutes knowledge. The first clause in Definition 2.20 allows us to form *complex* bases out of atomic ones. If b is a basis for the agent's believing p and b^* is a basis for the agent's believing $p \to q$, then $b^* \cdot b$ is "the basis for believing q which consists of believing p on basis b, believing $p \to q$ on basis b^*, and believing q by competent deduction from those premises" [Williamson, 2009, 20].[62] For methods instead of bases, Nozick says that if one comes to believe p by method M_1 and $p \to q$ by method M_2, then by deducing q from p and $p \to q$, one comes to believe q by the "combined" method "$\mathrm{M}_1 + \mathrm{M}_2$" [Nozick, 1981, 233], which we would write as $\mathrm{M}_2 \cdot \mathrm{M}_1$. Although I do not endorse this way of thinking about deduction, I will go along with Williamson and Nozick in order to study their views on their own terms.

We interpret the basis language in models that enrich CB models with a function $\mathcal{B}$, inspired by the function $\mathcal{E}$ of Fitting models for justification logic (see Artemov 2008, §5, Fitting 2009, §4.2) but with different properties.

Definition 2.21 (CB* Model)**.** A CB* model is a tuple $\mathcal{M} = \langle W, D, \mathcal{B}, \leqslant, V\rangle$ where W, D, $\leqslant$, and V are as in Definition 2.2 and $\mathcal{B}$ assigns to each pair of a basis/method b (atomic or complex) and φ in the language a set $\mathcal{B}(\varphi, b) \subseteq W$ such that:

1. if $\mathcal{B}(\varphi, b) \neq \emptyset$, then $\mathcal{B}(\psi, b) = \emptyset$ for $\psi \neq \varphi$;

2. if $\mathcal{B}(\varphi \to \psi, b) \neq \emptyset$ and $\mathcal{B}(\varphi, b') \neq \emptyset$, then $\mathcal{B}(\psi, b \cdot b') \subseteq \mathcal{B}(\varphi \to \psi, b) \cap \mathcal{B}(\varphi, b')$.

We interpret $\mathcal{B}(\varphi, b)$ to be the set of worlds in which b is a basis for believing φ (see Definition 2.23 below). Parts 1 and 2 of the definition follow from our interpretation of bases/methods along the lines of Williamson/Nozick above. First, we allow there to be multiple bases for the same belief. We might also wish to say that an agent's basis for believing φ is the same as her basis for believing ψ. However, we will adopt the convention that bases are to be individuated in part by what they are bases for: if b is a basis for believing φ in some world, then b is not a basis for believing some other ψ in any world. (Similarly, we individuate methods in part by the belief content

[62] Harman and Sherman [2004] deny Williamson's "presupposition that deduction is a kind of inference, something one does," arguing that this "confuses questions of implication with questions of inference" (495). I will return to this issue in §5.4.

they can produce.) The reason for this convention, imposed by Definition 2.21.1, is that if we understand the basis $b^* \cdot b$ for believing q as in the quote from Williamson, then $b^* \cdot b$ cannot be a basis for believing some other r. Definition 2.21.2 also follows from understanding bases of the form $b \cdot b'$ following Williamson. If b is a basis for believing $\varphi \to \psi$ and b' is a basis for believing φ, then in any world where the agent believes ψ on the basis of $b \cdot b'$, she must also believe $\varphi \to \psi$ on the basis of b and φ on the basis of b'. Similar points apply to methods, following the passage from Nozick.

We now define H*- and S*-semantics as modifications of H- and S-semantics. As noted above, we ignore adherence and N-semantics for simplicity.

Definition 2.22 (Truth in CB* Models)**.** Given a well-founded CB* model $\mathcal{M} = \langle W, D, \mathcal{B}, \leqslant, V\rangle$ with $w \in W$ and a formula φ in the basis language, define $\mathcal{M}, w \vDash_x \varphi$ as follows (with propositional cases as usual):

$\mathcal{M}, w \vDash_x B(\varphi, b)$ iff $w \in \mathcal{B}(\varphi, b)$ and $\forall v \in W$: if wDv then $\mathcal{M}, v \vDash \varphi$;

$\mathcal{M}, w \vDash_{h^*} K(\varphi, m)$ iff $\mathcal{M}, w \vDash_{h^*} B(\varphi, m)$ and
(sensitivity*) $\forall v \in \mathrm{Min}_{\leqslant_w}\big(\overline{[\![\varphi]\!]}_{h^*}\big)$: $\mathcal{M}, v \nvDash_{h^*} B(\varphi, m)$;

$\mathcal{M}, w \vDash_{s^*} K(\varphi, b)$ iff $\mathcal{M}, w \vDash_{s^*} B(\varphi, b)$ and
(safety*) $\forall v \in \mathrm{Min}_{\leqslant_w}(W)$: $\mathcal{M}, v \vDash_{s^*} B(\varphi, b) \to \varphi$.

These clauses clearly capture the sensitivity* and safety* conditions stated at the beginning of this section. Note that I have incorporated Observation 2.1 into the S*-clause (and the remarks of 2.B apply here as well). There are different options for defining the truth of $B\varphi$ and $K\varphi$ formulas,[63] but we will not need them here.

[63]For $K\varphi$ we could define $\mathcal{M}, w \vDash_x K\varphi$ iff there exists a $b \in \mathsf{Ba}$ such that $\mathcal{M}, w \vDash_x K(\varphi, b)$. According to Nozick [1981, 181], it is not enough for there to be one method with respect to which the agent's belief is sensitive*. Nozick proposes a more subtle analysis in terms of which methods *outweigh* other methods, while Luper-Foy [1984, 27n4] suggests that one good method should be enough. In any case, this issue is not essential to our closure question. For $B\varphi$, we could adopt the clause in Definition 2.7, or we could define $\mathcal{M}, w \vDash B\varphi$ iff there exists a $b \in \mathsf{Ba}$ such that $\mathcal{M}, w \vDash_x B(\varphi, b)$. Note that while the $B\varphi$ clause in Definition 2.7 guarantees full *doxastic closure*, the clause just given does not. The latter allows $B\varphi \wedge \neg B\psi$ to be satisfiable even when ψ is a logical consequence of φ. We could think of the clause in Definition 2.7 as defining "implicit" belief, while

Given the conditions on CB* models we have imposed so far, we can have both $w \in \mathcal{B}(\varphi, b)$ and $\mathcal{M}, w \not\models B(\varphi, b)$. In other words, the models allow that b is a basis for believing φ even though the agent does not believe φ on that basis. However, we can also choose to represent a basis in our model only if the agent believes something on that basis. As I will put it, we can represent only "realized bases" in the model.

Definition 2.23 (Realized Bases)**.** A CB* model has *realized bases* iff for all $w \in W$, $b \in \mathsf{Ba}$, and formulas φ in the basis language,

$$w \in \mathcal{B}(\varphi, b) \text{ implies } \mathcal{M}, w \models B(\varphi, b).$$

With this setup, let us now turn to the issue of deductive closure for the safety* theory. Fact 2.15 and its proof formalize an argument, clearly stated by Williamson [2009, 20], according to which safety* preserves deductive closure.

Fact 2.15 (Deductive Closure & Bases)**.** The deductive closure principle

$$(K(\varphi \to \psi, b) \wedge K(\varphi, b') \wedge B(\psi, b \cdot b')) \to K(\psi, b \cdot b')$$

is S*-valid over CB* models with realized bases.

Proof. Assume

$$\mathcal{M}, w \models_{s^*} K(\varphi \to \psi, b) \wedge K(\varphi, b') \wedge B(\psi, b \cdot b'). \tag{2.67}$$

It follows by Definition 2.22 that $\mathcal{B}(\varphi \to \psi, b) \neq \emptyset$ and $\mathcal{B}(\varphi, b') \neq \emptyset$, which with Definition 2.21.2 implies

$$\mathcal{B}(\psi, b \cdot b') \subseteq \mathcal{B}(\varphi \to \psi, b) \cap \mathcal{B}(\varphi, b'). \tag{2.68}$$

we interpret $B(\varphi, b)$ to mean that the agent has an "explicit" (though not necessarily occurrent) belief that φ on the basis of b. This interpretation makes sense of the satisfiability of $B\varphi \wedge \neg B\psi$ according to the clause just given, even when ψ is a consequence of φ. For the agent may not have formed an explicit belief that ψ. Rather than choosing between the $B\varphi$ clause of Definition 2.7 and the other clause, we could introduce different belief operators for the different clauses.

Consider some $v \in \mathrm{Min}_{\leqslant_w}(W)$. If $\mathcal{M}, v \vDash_{s^*} B(\psi, b \cdot b')$, then by Definition 2.22, (2.68), and Definition 2.23,

$$\mathcal{M}, v \vDash_{s^*} B(\varphi \to \psi, b) \wedge B(\varphi, b'). \tag{2.69}$$

By Definition 2.22, (2.69) and (2.67) together imply $\mathcal{M}, v \vDash_{s^*} (\varphi \to \psi) \wedge \varphi$ and hence $\mathcal{M}, v \vDash_{s^*} \psi$. We conclude that for all $v \in \mathrm{Min}_{\leqslant_w}(W)$, $\mathcal{M}, v \vDash_{s^*} B(\psi, b \cdot b') \to \psi$, which with $\mathcal{M}, w \vDash_{s^*} B(\psi, b \cdot b')$ implies $\mathcal{M}, w \vDash_{s^*} K(\psi, b \cdot b')$ by Definition 2.22. □

The key to Fact 2.15 is Williamson's assumption (similar to Nozick's for methods) that the basis for believing ψ is not just *deduction from $\varphi \to \psi$ and φ, which are believed*, but rather *deduction from $\varphi \to \psi$ and φ, which are believed on bases b and b', respectively*. Alspector-Kelly [2011] objects that building the bases for the premises into the basis for the conclusion so that safety* saves deductive closure is *ad hoc*. I will not enter this dispute, since my main objection to safety/safety* is not about deductive closure, but about the Problem of Vacuous Knowledge explained in §4.1.

The second point about Fact 2.15 is that if safety* saves closure at all, it only saves *deductive* closure. As suggested by Williamson [2000, 282f], it is highly plausible that an agent who knows $\varphi \wedge \psi$ thereby knows ψ, in virtue of knowing the conjunction, even if the agent did not *deduce* ψ from $\varphi \wedge \psi$ (recall Remark 2.1). Yet the safety* theory fails to deliver this result. Recall the CB model in Fig. 2.10. At w, the agent believes p and q, and she safely* believes $p \wedge q$, since it is true at all of the closest worlds to w where she believes $p \wedge q$ on the same basis as she does in w (which is only w itself in Fig. 2.10, but we could add more). Assume that in w her belief that q is not based on deduction from $p \wedge q$, but (solely) on some other basis b, which does not depend on her believing p or $p \wedge q$. Then we may assume that b is also her basis for believing q in v, a world in which she believes neither p nor $p \wedge q$. Since q is false at v, which is a closest (to w) world, she does not have a safe basis for believing q at w. Hence according to the safety* theory, an agent can know $p \wedge q$ and believe q, but fail to know q, contradicting the view suggested by Williamson. I take this to be a problem for the safety* theory, rather than for Williamson's suggestion.

I conclude this section by noting that even if we allow Nozick's conception of combined methods, sensitivity* does not save deductive closure.

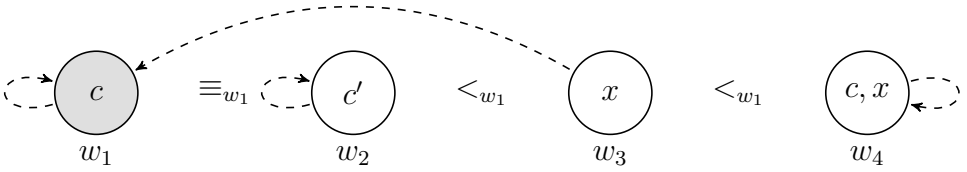

Figure 2.14: CB model for Example 1.1 (partially drawn)

Fact 2.16 (Deductive Closure & Methods)**.** The deductive closure principle

$$(K(\varphi \to \psi, m) \wedge K(\varphi, m') \wedge B(\psi, m \cdot m')) \to K(\psi, m \cdot m') \tag{2.70}$$

is not H*-valid over CB* models (even with realized bases/methods).

For the proof, we return to Example 1.1 and the model in Fig. 2.14.

Proof. Let us extend the CB model $\mathcal{M} = \langle W, D, \leqslant, V\rangle$ in Fig. 2.14 to a CB* model $\mathcal{M}^* = \langle W, D, \mathcal{B}, \leqslant, V\rangle$ by defining $\mathcal{B}$ in any way such that

$$\{w_1, w_3\} \subseteq \mathcal{B}(c \to \neg x, m) \cap \mathcal{B}(c, m') \cap \mathcal{B}(\neg x, m \cdot m').^{64} \tag{2.71}$$

In (2.70), substitute c for φ and $\neg x$ for ψ to obtain

$$(K(c \to \neg x, m) \wedge K(c, m') \wedge B(\neg x, m \cdot m')) \to K(\neg x, m \cdot m'). \tag{2.72}$$

Given $\mathrm{Min}_{\leqslant_{w_1}}\big(\overline{[\![c \to \neg x]\!]}\big) = \{w_4\}$ and $\mathcal{M}, w_4 \nvDash B(c \to \neg x, m)$ (since $w_4 D w_4$ and $\mathcal{M}, w_4 \nvDash c \to \neg x$), sensitivity* holds at w_1 for $c \to \neg x$ with m. Then since $\mathcal{M}, w_1 \vDash B(c \to \neg x, m)$, it follows that $\mathcal{M}, w_1 \vDash_{h^*} K(c \to \neg x, m)$. Similarly, given $\mathrm{Min}_{\leqslant_w}\big(\overline{[\![c]\!]}\big) = \{w_2\}$ and $\mathcal{M}, w_2 \nvDash B(c, m')$ (since $w_2 D w_2$ and $\mathcal{M}, w_2 \nvDash c$), sensitivity* holds at w_1 for c with m'. Then since $\mathcal{M}, w \vDash B(c, m')$, it follows that $\mathcal{M}, w_1 \vDash_{h^*} K(c, m')$. But given $\mathrm{Min}_{\leqslant_{w_1}}\big(\overline{[\![\neg x]\!]}\big) = \{w_3\}$ and $\mathcal{M}, w_3 \vDash B(\neg x, m \cdot m')$,

[64] Think of m as the method of looking up in the medical textbooks whether $c \to \neg x$ and m' as the method of running the laboratory tests to check for signs of condition c. Assume that in both w_1, the (actual) world in which the patient has condition c, and w_3, the world in which the patient has disease x, student A forms her belief in $c \to \neg x$ by method m, her belief in c by method m', and her belief in $\neg x$ by deduction from $c \to \neg x$ and c.

sensitivity* fails at w_1 for $\neg x$ with $m \cdot m'$, so $\mathcal{M}, w_1 \not\models_{h^*} K(\neg x, m \cdot m')$. Then since $\mathcal{M}, w_1 \models B(\neg x, m \cdot m')$, it follows that the conditional (2.72) is false at $\mathcal{M}, w_1$. □

The difference between Fact 2.15 and Fact 2.16 is due to the $\exists\forall$ nature of safety* vs. the $\forall\exists$ nature of sensitivity*. A world v that witnesses the violation of either sensitivity* or safety* at w with respect to ψ must be a $\neg\psi$-world and hence a $\neg\varphi$-world or a $\neg(\varphi \to \psi)$-world. Since sensitivity* is a $\forall\exists$ condition, if the $\neg\psi$-world v witnesses the violation of sensitivity* at w with respect to ψ, it need not be among the worlds that matter for sensitivity* at w with respect to φ and $\varphi \to \psi$, namely the *closest* (to w) $\neg\varphi$-worlds or $\neg(\varphi \to \psi)$-worlds. Hence even if in v the agent comes to (falsely) believe φ or $\varphi \to \psi$ by the same methods as in w,[65] the agent may still sensitively* believe φ and $\varphi \to \psi$ at w. By contrast, since safety is a $\exists\forall$ condition, if the $\neg\psi$-world v witnesses the violation of safety* at w with respect to ψ, it must be among worlds that matter for safety* at w with respect to φ and $\varphi \to \psi$, namely the closest worlds to w. Hence if in v the agent (falsely) believes φ or $\varphi \to \psi$ on the same basis as in w, then the agent cannot safely* believe φ and $\varphi \to \psi$ at w.

The $\forall\exists$ nature of sensitivity* also explains the "complication" in Nozick's [1981, 236] explanation for why closure under "inferring a disjunction from a disjunct" should hold for his theory when methods are taken into account. Suppose that in w, the agent knows p and infers $p \vee q$ from p. Further suppose that v is a closest (to w) $\neg(p \vee q)$-world, and therefore a $\neg p$-world. Nozick suggests that in v, the agent cannot believe p—and thus cannot come to believe $p \vee q$ by the same method of inference from p, as used in w—because if she does believe p at v, then at w her belief that p is not sensitive. But this assumes that v, which is a closest (to w) $\neg(p \vee q)$-world, is also a closest (to w) $\neg p$-world, an assumption there is no reason to make. The closest $\neg p$-worlds may all be q-worlds. Therefore, Nozick's explanation is incorrect.

[65] One might claim that in w, the method used is: deduction from φ and $\varphi \to \psi$, which are *truly* believed by methods m and m'. If so, this cannot be the method used in v, since either φ or $\varphi \to \psi$ must be false at v. One might claim that in v, the *different* method used is: deduction from φ and $\varphi \to \psi$, which are (merely) believed by methods m and m'. However, Nozick [1981, 232] himself rejects this suggestion, since the "two methods" described are *indistinguishable for the agent*, and Nozick individuates methods "from the inside" (184).

2.E Subjunctivist vs. Probabilistic Models

As noted in §2.5, Roush [2005] argues that sensitivity and adherence should be understood in terms of conditional probability rather than subjunctive conditionals. Where s and t are parameters, sensitivity becomes $P(\neg Bp \mid \neg p) > s$, adherence becomes $P(Bp \mid p) > t$,[66] and an agent *Nozick-knows* p (relative to s, t) iff p is true, the agent believes p, and sensitivity and adherence are met relative to s and t, respectively. Roush's own view of knowledge, sketched in §2.C, has another disjunctive clause.

In this section, I show that the probabilistic version of Nozick-knowledge (to be distinguished from Roush's more sophisticated account of knowledge) leads to the same closure failures as the subjunctivist version. To do so, I will draw on results of Baltag and Smets [2008] on the relationship between preorders and "Popper functions" for conditional probability. Instead of starting with a probability function $P : \mathcal{P}(W) \to [0, 1]$ on our space W, from which conditional probability can be defined, we will take conditional probability as primitive. Baltag and Smets treat the case where W is finite, which will be sufficient for our purposes here.

Definition 2.24 (Discrete Popper Function)**.** A *discrete Popper function* on a finite set W is a function $\mu\colon \mathcal{P}(W) \times \mathcal{P}(W) \to [0, 1]$ such that for all $A, B, C \subseteq W$:

1. $\mu(A \mid A) = 1$;
2. $\mu(A \cup B \mid C) = \mu(A \mid C) + \mu(B \mid C)$, if $A \cap B = \emptyset$ and $C \neq \emptyset$;
3. $\mu(A \cap B \mid C) = \mu(A \mid B \cap C) \cdot \mu(B \mid C)$.

An absolute probability function $P : \mathcal{P}(W) \to [0, 1]$ can be defined from the Popper function μ by

$$P(A) = \mu(A \mid W).$$

Conditional probability is usually defined in terms of absolute probability by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)},$$

[66] Roush also include $P(B\neg p \mid p) < 1 - t$ as a conjunct in the adherence condition, since she does not assume that the agent is fully rational according to the probability axioms.

so note that it follows from Definition 2.24.3 that if $P(B) \neq 0$, then

$$\frac{P(A \cap B)}{P(B)} = \frac{\mu(A \cap B \mid W)}{\mu(B \mid W)} = \mu(A \mid B \cap W) = \mu(A \mid B),$$

so the derived conditional probability agrees with the primitive conditional probability. Also note that even if $P(B) = 0$, $\mu(A \mid B)$ is well-defined, which is to say that Popper functions allow conditionalization on propositions of probability 0.

2.E.1 Probabilistic Tracking (PT) Models

Let us now formalize probabilistic Nozick knowledge with new models and semantics.

Definition 2.25 (PT Model)**.** A *probabilistic tracking model* is a tuple $\mathfrak{M} = \langle W, D, \mu, V \rangle$ where W is a finite set, D and V are defined as in Definition 2.6, and

1. μ assigns to each $w \in W$ a function $\mu_w \colon \mathcal{P}(W) \times \mathcal{P}(W) \to [0, 1]$;
   (a) μ_w is a Popper function;
   (b) $\mu_w(\{w\} \mid W) > 0$.

For the following definition, fix some thresholds $s, t \in [0, 1)$.

Definition 2.26 (Truth in a PT Model)**.** Given a PT model $\mathfrak{M} = \langle W, D, \mu, V \rangle$ with $w \in W$ and φ in the epistemic-doxastic language, define $\mathfrak{M}, w \Vdash \varphi$ as follows (with propositional cases as usual):

$$\begin{aligned} \mathfrak{M}, w \Vdash K\varphi \quad \text{iff} \quad & \mathfrak{M}, w \Vdash B\varphi \wedge \varphi \text{ and} \\ & \text{(sensitivity) } \mu_w(\overline{\|B\varphi\|} \mid \overline{\|\varphi\|}) > s, \\ & \text{(adherence) } \mu_w(\|B\varphi\| \mid \|\varphi\|) > t, \end{aligned}$$

where $\|\alpha\| = \{v \in W \mid \mathcal{M}, v \vDash \alpha\}$. Observe that the condition that $\mathfrak{M}, w \Vdash \varphi$ is not redundant. However, given Definition 2.25.1b, the condition that $\mathfrak{M}, w \Vdash \varphi$ would be redundant if we were to define sensitivity as $\mu_w(\overline{\|B\varphi\|} \mid \overline{\|\varphi\|}) = 1$; moreover, if we were to assume that $\mu_w(\{w\} \mid W) = 1$ (analogous to the *centering* condition

that $\mathrm{Min}_{\leqslant_w}(W) = \{w\}$) instead of just $\mu_w(\{w\} \mid W) > 0$, then the condition that $\mathfrak{M}, w \Vdash \varphi$ would be redundant even with sensitivity as defined above.

2.E.2 From CB to PT Models

To connect PT models to CB models, we define Popper functions μ_w from the preorders $\leqslant_w$ in our CB models in two steps, using the following definition and theorem.

Given a preorder $\preceq$ on a finite set W, define the function $(\cdot\,,\cdot)_{\preceq} : W \times W \to [0,1]$ as follows:

$$(w,v)_{\preceq} = \begin{cases} 1 \text{ if } w \prec v \text{ or } w = v; \\ 0 \text{ if } v \prec w; \\ .5 \text{ otherwise.} \end{cases}$$

Theorem 2.2 (Baltag and Smets 2008)**.** If $\preceq$ is a preorder on a finite set W, then the function $\mu\colon \mathcal{P}(W) \times \mathcal{P}(W) \to [0,1]$ defined by[67]

$$\mu(A \mid B) = \sum_{w \in A \cap B} \frac{1}{\sum\limits_{v \in B} \frac{(v,w)_{\preceq}}{(w,v)_{\preceq}}}$$

is a Popper function such that for all $A, B \subseteq W$,

$$\mu(A \mid B) = 1 \text{ iff } \mathrm{Min}_{\preceq}(B) \subseteq A. \tag{2.73}$$

Finally, using Theorem 2.2 we can show that closure fails for probabilistic Nozick knowledge in all of the ways that it fails for subjunctivist Nozick knowledge.

Proposition 2.9 (Closure Failures from CB to PT Models)**.** Any flat closure principle that is not H-valid over CB models is not valid over PT models.

Proof. It follows from the proof of Theorem 2.1 that if a flat closure principle

$$K\varphi_1 \wedge \cdots \wedge K\varphi_n \to K\psi_1 \vee \cdots \vee K\psi_m$$

[67] Using the conventions that $\frac{1}{0} = \infty$, $\frac{1}{\infty} = 0$, $\infty + \infty = \infty$, and $\infty + x = \infty$ for all $x \in \mathbb{R}$.

is not N-valid over CB models, then it is falsified at a finite pointed CB model $\mathcal{M}, w$ where for all $j \leq m$, not only

$$\mathrm{Min}_{\leqslant_w}(\overline{[\![\psi_j]\!]}) \not\subseteq \overline{[\![B\psi_j]\!]},$$

but also

$$\mathrm{Min}_{\leqslant_w}(\overline{[\![\psi_j]\!]}) \subseteq [\![B\psi_j]\!].^{68} \tag{2.74}$$

From the CB model $\mathcal{M} = \langle W, D, \leqslant, V\rangle$, define the PT model $\mathfrak{M} = \langle W, D, \mu, V\rangle$ by constructing μ_w from $\leqslant_w$ as in Theorem 2.2.

From (2.74), it follows by (2.73) that

$$\mu_w(\|B\psi_j\| \mid \overline{\|\psi_j\|}) = 1,$$

in which case

$$\mu_w(\overline{\|B\psi_j\|} \mid \overline{\|\psi_j\|}) = 0$$

by Definition 2.24.2. It follows that no matter the value of s, we have $\mathfrak{M}, w \nVdash K\psi_j$ by Definition 2.26.

For all $i \leq n$, given $\mathcal{M}, w \vDash K\varphi_i$ we have

$$\mathrm{Min}_{\leqslant_w}(\overline{[\![\varphi_i]\!]}) \subseteq \overline{[\![B\varphi_i]\!]}$$

and

$$\mathrm{Min}_{\leqslant_w}([\![\varphi_i]\!]) \subseteq [\![B\varphi_i]\!]$$

by Definition 2.7, so

$$\mu_w(\overline{\|B\varphi_i\|} \mid \overline{\|\varphi_i\|}) = 1$$

[68] There are multiple ways to construct such a model. For one, use the proof of Theorem 2.1.2 to first construct a *linear* RA model that falsifies the closure principle according to D-semantics; then use Proposition 2.4/2.5 to obtain a linear CB model that falsifies it according to N-semantics. In a linear model, $\mathrm{Min}_{\leqslant_w}(\overline{[\![\psi]\!]}) \not\subseteq \overline{[\![B\psi]\!]}$ obviously implies $\mathrm{Min}_{\leqslant_w}(\overline{[\![\psi]\!]}) \subseteq [\![B\psi]\!]$. For another, use the construction of §2.6.4. To be more concrete, observe that in the model $\mathcal{M}$ in Fig. 2.2, we have $\mathcal{M}, w_1 \nvDash_n K(c \wedge \neg x) \rightarrow K\neg x$ and $\mathrm{Min}_{\leqslant_{w_1}}(\overline{[\![\neg x]\!]}) = \{w_3\} \subseteq [\![B\neg x]\!]$.

and

$$\mu_w(\|B\varphi_i\|| \, \|\varphi_i\|) = 1$$

by (2.73). Also note that $\mathcal{M}, w \vDash K\varphi_i$ implies $\mathcal{M}, w \vDash B\varphi_i \wedge \varphi_i$, which implies $\mathfrak{M}, w \vDash B\varphi_i \wedge \varphi_i$ since φ_i is propositional and V and D are the same in $\mathcal{M}$ and $\mathfrak{M}$. It follows that no matter the value of t, we have $\mathfrak{M}, w \Vdash K\varphi_i$ by Definition 2.26. Putting this together with our results from above, we have

$$\mathfrak{M}, w \nVdash K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi_1 \vee \cdots \vee \psi_m$$

no matter the values of s and t. □

Is the converse of Proposition 2.9 true, so that all principles valid over CB models in N-semantics are also valid over PT models? The answer is negative:

Fact 2.17 (C Axiom)**.** For any $s, t \in [0, 1)$, the C axiom $K\varphi \wedge K\psi \rightarrow K(\varphi \wedge \psi)$ is not valid over PT models.

Proof. It is easy to construct a pointed PT model $\mathfrak{M}, w$ where

$$\mathfrak{M}, w \vDash Bp \wedge Bq \wedge p \wedge q, \tag{2.75}$$

$$\mu_w(\overline{\|Bp\|} \mid \overline{\|p\|}) = 1, \tag{2.76}$$

$$\mu_w(\overline{\|Bq\|} \mid \overline{\|q\|}) = 1, \tag{2.77}$$

$$\mu_w(\|Bp\| \mid \|p\|) > t, \text{ and} \tag{2.78}$$

$$\mu_w(\|Bq\| \mid \|q\|) > t, \tag{2.79}$$

but where

$$\mu_w(\|B(p \wedge q)\| \mid \|p \wedge q\|) \leq t. \tag{2.80}$$

By (2.75) - (2.79), the belief, truth, sensitivity, and adherence conditions are satisfied for p and q at w, but by (2.80), adherence is not satisfied for $p \wedge q$ at w.

The C axiom can also be falsified due to a lack of sensitivity to $p \wedge q$, rather than a

lack of adherence. For $s > 0$, it is easy to construct a pointed PT model $\mathfrak{M}, w$ where

$$\mu_w(\overline{\|Bp\|} \mid \overline{\|p\|}) > s, \tag{2.81}$$

$$\mu_w(\overline{\|Bq\|} \mid \overline{\|q\|}) > s, \tag{2.82}$$

and (2.75), (2.78), and (2.79) hold, but where

$$\mu_w(\overline{\|B(p \wedge q)\|} \mid \overline{\|p \wedge q\|}) < s, \tag{2.83}$$

so sensitivity is not satisfied for $p \wedge q$ at w.[69] □

2.F Reduction Axioms for RA Context Change

In this section, our goal is to apply one of the main ideas of dynamic epistemic logic, that of *reduction axioms*, to the picture of context change presented in §2.11. Roughly speaking, reduction axioms are valid equivalences of the form $[+\chi]\psi \leftrightarrow \psi'$, where the left-hand side states that some ψ is true *after* the context change with χ, while the right-hand side gives an *equivalent* ψ' describing what is true *before* the context change. For example, we can ask whether an agent counts as knowing φ after χ becomes relevant, i.e., is $[+\chi]K\varphi$ true? The reduction axioms will answer this question by describing what must be true of the agent's epistemic state *before* the context change in order for the agent to count as knowing φ after the context change.

To obtain reduction axioms for context change that are valid over our RA models, we will use a language more expressive than the epistemic language used in the previous sections. Our new *RA language* will be capable of describing what is relevant at a world and what is ruled out at a world independently. This additional expressive power will allow us to obtain reduction axioms using methods similar to those applied

[69] For example, construct $\mathfrak{M}$ with worlds w, x_1, x_2, and y, where $\mathfrak{M}, w \Vdash B(p \wedge q) \wedge p \wedge q$, $\mathfrak{M}, x_1 \Vdash B(p \wedge q) \wedge p \wedge \neg q$, $\mathfrak{M}, x_2 \Vdash B(p \wedge q) \wedge \neg p \wedge q$, and $\mathfrak{M}, y \Vdash \neg Bp \wedge \neg Bq \wedge \neg p \wedge \neg p$. For $s \in (0, 1)$, let the unconditional probabilities of the singleton sets relative to w be: $P_w(\{w\}) = k$ for some $k \in (0, 1)$, $P_w(\{x_1\}) = P_w(\{x_2\}) = \frac{2(1-k)(1-s)}{(4-s)}$, and $P_w(\{y\}) = \frac{3sP_w(\{x_i\})}{2(1-s)}$, which sum to 1. Then one can check that $P_w(\overline{\|Bp\|} \mid \overline{\|p\|}) = P_w(\overline{\|Bq\|} \mid \overline{\|q\|}) > s$, but $P_w(\overline{\|B(p \wedge q)\|} \mid \overline{\|p \wedge q\|}) < s$.

by van Benthem and Liu van Benthem and Liu [2007] to *dynamic epistemic preference logic* (also see van Benthem et al. [2009]), but with an important difference.

Van Benthem and Liu work with models with a *single* preorder over worlds (for each agent), representing an agent's preferences between worlds, and their language contains an operator $\Box^{\succ}$ used to quantify over all worlds that are *better* than the current world according to the agent.[70] In our setting, $\Box^{\succ}$ would quantify over all worlds that are *more relevant*. Using another operator $\Box^{\twoheadrightarrow}$ to quantify over all worlds that are *uneliminated* at the current world, we can try to write a formula expressing that all of the most relevant $\neg\varphi$-worlds are eliminated at the current world. An equivalent statement is that for all uneliminated worlds v, if v is a $\neg\varphi$-world, then there is another $\neg\varphi$-world that is strictly more relevant than v. This is expressed by $\Box^{\twoheadrightarrow}(\neg\varphi \rightarrow \Diamond^{\succ}\neg\varphi)$, where $\Diamond^{\succ}\psi := \neg\Box^{\succ}\neg\psi$.

The problem with the above approach is that unlike the models of van Benthem and Liu (but like models for conditional logic and the general *belief revision structures* of Board [2004]), our RA models include a preorder $\preceq_w$ for *each* world w. Hence if the operator $\Box^{\succ}$ quantifies over all worlds that are more relevant than the current world according to the relevance relation of the current world, then $\Box^{\twoheadrightarrow}(\neg\varphi \rightarrow \Diamond^{\succ}\neg\varphi)$ will be true at w just in case for all worlds v uneliminated at w, if v is a $\neg\varphi$-world, then there is another $\neg\varphi$-world that is strictly more relevant than v *according to* $\preceq_v$. Yet this is not the desired truth condition.[71] The desired truth condition is that for all worlds v uneliminated at w, if v is a $\neg\varphi$-world, then there is another $\neg\varphi$-world that is strictly more relevant than v *according to* $\preceq_w$. To capture this truth condition, we will use an approach inspired by *hybrid logic* [Areces and ten Cate, 2007]. First, different modalities $\Box^{\succ_x}$, $\Box^{\succ_y}$, etc., will be associated in a given model with different relevance relations $\preceq_w$, $\preceq_v$, etc., by an assignment function g. Second, a *binder* $\downarrow$ will be used to bind a world variable x to the current world, so that the formula

[70] Van Benthem et al. [2009] write this operator as $\Box^{<}$, since they take $w \prec v$ to mean that v is strictly better than w according to the agent. Since we take $w \prec v$ to mean that w is strictly more relevant than v, we write $\Box^{\succ}$ for the operator that quantifies over more relevant worlds. We will write $\Box^{\preceq}$ for the operator that quantifies over worlds that are of equal or lesser relevance. We use the same $\preceq$ for the superscript of the operator and for the relation in the model, trusting that no confusion will arise.

[71] Since v is assumed to be minimal in $\preceq_v$, the condition would never be met.

$\downarrow x.\Box^{\rightarrow}(\neg\varphi \rightarrow \Diamond^{\succ_x}\neg\varphi)$ will capture the desired truth condition described above (cf. Lewis 1973, §2.8 on the † operator).

In addition to the operator $\Box^{\succ_x}$ that quantifies over all worlds more relevant than the current world according to $\preceq_{g(x)}$, we will use an operator $\Box^{\preceq_x}$ that quantifiers over all worlds whose relevance is equal to or lesser than that of the current world according to $\preceq_{g(x)}$. The second operator is necessary for writing reduction axioms for the context change operation $\nwarrow$ from Definition 2.15. Together the two types of operators will also allow us to quantify over all worlds in the field of $\preceq_{g(x)}$, $W_{g(x)}$, with formulas of the form $\Box^{\succ_x}\varphi \wedge \Box^{\preceq_x}\varphi$, which we will use in writing reduction axioms for both of the context change operations, $\uparrow$ and $\nwarrow$, from Definition 2.15.

Definition 2.27 (Dynamic & Static RA Languages)**.** Let $\mathsf{At} = \{p, q, r \dots\}$ be a set of atomic sentences and $\mathsf{Var} = \{x, y, z, \dots\}$ a set of variables. The *dynamic RA language* is generated as follows, where $p \in \mathsf{At}$ and $x \in \mathsf{Var}$:

$$\begin{aligned} \varphi &::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \Box^{\rightarrow}\varphi \mid \Box^{\preceq_x}\varphi \mid \Box^{\succ_x}\varphi \mid \downarrow x.\varphi \mid [\pi]\varphi \\ \pi &::= \uparrow \varphi \mid \nwarrow \varphi. \end{aligned}$$

Where R is $\preceq_x$, $\succ_x$, or $\rightarrow$, let $\Diamond^R\varphi := \neg\Box^R\neg\varphi$; let R_x stand for either $\preceq_x$ or $\succ_x$ in definitions that apply to both; and let us use $+$ as after Definition 2.15. Finally, let the *static RA language* be the fragment of the dynamic RA language consisting of those formulas that do not contain any context change operators $[\pi]$.

The truth clauses are as one would expect from our description above, and the clause for the context change operators is the same as Definition 2.17.

Definition 2.28 (Truth)**.** Given an RA model $\mathcal{M} = \langle W, \rightarrow, \preceq, V\rangle$ and an assignment function $g\colon \mathsf{Var} \rightarrow W$, we define $\mathcal{M}, g, w \vDash \varphi$ as follows (with propositional cases as in Definition 2.4):

$$\begin{aligned} \mathcal{M}, g, w \vDash \Box^{\rightarrow}\varphi \quad &\text{iff} \quad \forall v \in W\colon \text{if } w \rightarrow v \text{ then } \mathcal{M}, g, v \vDash \varphi; \\ \mathcal{M}, g, w \vDash \Box^{R_x}\varphi \quad &\text{iff} \quad \forall v \in W\colon \text{if } wR_{g(x)}v \text{ then } \mathcal{M}, g, v \vDash \varphi; \\ \mathcal{M}, g, w \vDash [+\chi]\varphi \quad &\text{iff} \quad \mathcal{M}^{+\chi}, g, w \vDash \varphi; \\ \mathcal{M}, g, w \vDash \downarrow x.\varphi \quad &\text{iff} \quad \mathcal{M}, g^x_w, w \vDash \varphi, \end{aligned}$$

where g^x_w is such that $g^x_w(x) = w$ and $g^x_w(y) = g(y)$ for all $y \neq x$.

Hence the $\downarrow x.\varphi$ clause captures the idea of letting x stand for the current world by changing the assignment g to one that maps x to w but is otherwise the same.

We now show how the epistemic language can be translated into the RA language in two different ways, corresponding to D- and L-semantics.[72] To simplify the translation, let us assume for the moment that all of our RA models $\mathcal{M} = \langle W, \rightarrow\!\!\!\!\triangleright, \preceq, V\rangle$ have the *universal field* property, so for all $w \in W$, $W_w = W$.

Definition 2.29 (Translation)**.** Let σ_d be the translation from the epistemic language of Definition 2.1 to the static RA language of Definition 2.27 defined by:

$$\begin{aligned}
\sigma_d(p) &= p \\
\sigma_d(\neg\varphi) &= \neg\sigma_d(\varphi) \\
\sigma_d(\varphi \wedge \psi) &= (\sigma_d(\varphi) \wedge \sigma_d(\psi)) \\
\sigma_d(K\varphi) &= \downarrow x.\Box^{\rightarrow\!\!\!\!\triangleright}(\neg\sigma_d(\varphi) \rightarrow \Diamond^{\succ_x}\neg\sigma_d(\varphi)).
\end{aligned}$$

Let σ_l be the translation with the same atomic and boolean clauses (with σ_l in place of σ_d) but with:

$$\sigma_l(K\varphi) = \downarrow x.\Box^{\rightarrow\!\!\!\!\triangleright}(\neg\sigma_l(\varphi) \rightarrow \Diamond^{\succ_x}\top).$$

As explained at the beginning of this section, the idea of the σ_d translation is that the truth clause for $K\varphi$ in D-semantics—stating that the most relevant $\neg\varphi$-worlds are eliminated—is equivalent to the statement that for all worlds v uneliminated at the current world w, if v is a $\neg\varphi$-world, then there is another $\neg\varphi$-world that is strictly more relevant than v according to $\preceq_w$. This is exactly what $\sigma_d(K\varphi)$ expresses. Similarly, the idea of the σ_l translation is that the truth clause for $K\varphi$ in L-semantics—stating that among the most relevant worlds overall, all $\neg\varphi$-worlds are eliminated—is equivalent to the statement that for all worlds v uneliminated at the current world w, if v is a $\neg\varphi$-world, then there is another world that is strictly more relevant than v according to $\preceq_w$, in which case v is not among the most relevant worlds overall according to

[72]Note that since the translation of Definition 2.29 only requires a single variable x, for our purposes here it would suffice to define the RA language such that $|\mathsf{Var}| = 1$.

$\preceq_w$. This is exactly what $\sigma_l(K\varphi)$ expresses. The following proposition confirms these claims.

Proposition 2.10 (Simulation)**.** For any RA model $\mathcal{M} = \langle W, \rightarrow\!\!\!\triangleright, \preceq, V\rangle$, assignment $g\colon \mathsf{Var} \to W$, world $w \in W$, and formula φ of the epistemic language:

$$\begin{aligned} \mathcal{M}, w \vDash_d \varphi \quad &\text{iff} \quad \mathcal{M}, g, w \vDash \sigma_d(\varphi); \\ \mathcal{M}, w \vDash_l \varphi \quad &\text{iff} \quad \mathcal{M}, g, w \vDash \sigma_l(\varphi). \end{aligned}$$

Proof. By induction on φ. All of the cases are trivial except where φ is of the form $K\psi$. By Definition 2.29, we are to show

$$\mathcal{M}, w \vDash_d K\psi \text{ iff } \mathcal{M}, g, w \vDash \downarrow x.\Box^{\rightarrow\!\!\!\triangleright}(\neg\sigma_d(\varphi) \to \Diamond^{\succ_x}\neg\sigma_d(\varphi)). \tag{2.84}$$

By Definition 2.28, the rhs of (2.84) holds iff for all $v \in W$, if $w \rightarrow\!\!\!\triangleright\, v$, then

$$\mathcal{M}, g^x_w, v \vDash \neg\sigma_d(\psi) \to \Diamond^{\succ_x}\neg\sigma_d(\psi). \tag{2.85}$$

By Definition 2.28, (2.85) is equivalent to the disjunction of the following:

$$\mathcal{M}, g^x_w, v \vDash \sigma_d(\psi); \tag{2.86}$$

$$\exists u \in W\colon u \prec_{g^x_w(x)} v \text{ and } \mathcal{M}, g^x_w, u \nvDash \sigma_d(\psi). \tag{2.87}$$

By the inductive hypothesis, (2.86) and (2.87) are respectively equivalent to

$$\mathcal{M}, v \vDash_d \psi \text{ and} \tag{2.88}$$

$$\exists u \in W\colon u \prec_w v \text{ and } \mathcal{M}, u \nvDash_d \psi. \tag{2.89}$$

Assuming $\mathcal{M}$ has the universal field property, the disjunction of (2.88) and (2.89) is equivalent to

$$v \notin \mathrm{Min}_{\preceq_w}(\overline{[\![\psi]\!]}). \tag{2.90}$$

Hence the rhs of (2.84) holds if and only if for all $v \in W$, if $w \rightarrow\!\!\!\triangleright\, v$, then (2.90) holds. The rhs of this biconditional is equivalent to the lhs of (2.84), $\mathcal{M}, w \vDash_d K\psi$,

by Definition 2.4. The proof for the case of L-semantics is similar. □

If we do not assume that RA models have the universal field property, then we must modify the translation of Definition 2.29 such that

$$\begin{aligned}\sigma'_d(K\varphi) &= \downarrow x.\Box^{\rightarrow}(\neg\sigma'_d(\varphi) \rightarrow (\Diamond^{\succ_x}\neg\sigma'_d(\varphi) \vee \Box^{\preceq_x}\bot));\\ \sigma'_l(K\varphi) &= \downarrow x.\Box^{\rightarrow}(\neg\sigma'_l(\varphi) \rightarrow (\Diamond^{\succ_x}\top \vee \Box^{\preceq_x}\bot)).\end{aligned}$$

We leave it to the reader to verify that given the modified translation, Proposition 2.10 holds without the assumption of the universal field property.

We are now prepared to do what we set out to do at the beginning of this section: give reduction axioms for the context change operations of Definition 2.15. For the following proposition, let us define $\Box^x\varphi := \Box^{\succ_x}\varphi \wedge \Box^{\preceq_x}\varphi$.

Proposition 2.11 (RA Reduction)**.** Given the valid reduction axioms in Table 2.3 below and the rule of replacement of logical equivalents,[73] any formula of the *dynamic* RA language is equivalent to a formula of the *static* RA language.

Proof. Assuming the axioms are valid, the argument for the claim of the proposition is straightforward. Each of the axioms drives the context change operators $[+\chi]$ inward until eventually these operators apply only to atomic sentences p, at which point they can be eliminated altogether using (2.91). In case we encounter something of the form $[+\chi_1][+\chi_2]\varphi$, we first reduce $[+\chi_2]\varphi$ to an equivalent static formula φ' and then use the replacement of logical equivalents to obtain $[+\chi_1]\varphi'$, which we then reduce to an equivalent static formula φ'', etc.

Let us now check the validity of (2.91) - (2.95) in turn. First, (2.91) is valid because the context change operations of Definition 2.15 do not change the valuation V for atomic sentences in the model. For (2.92), in the left-to-right direction we have the following implications: $\mathcal{M}, w \vDash [+\chi]\neg\varphi \Rightarrow \mathcal{M}^{+\chi}, w \vDash \neg\varphi \Rightarrow \mathcal{M}^{+\chi}, w \nvDash \varphi \Rightarrow \mathcal{M}, w \nvDash [+\chi]\varphi \Rightarrow \mathcal{M}, w \vDash \neg[+\chi]\varphi$. For the right-to-left direction of (2.92), simply

[73]Semantically, if $\alpha \leftrightarrow \beta$ is valid, so is $\varphi(\alpha/p) \leftrightarrow \varphi(\beta/p)$, where (ψ/p) indicates substitution of ψ for p.

$$
\begin{aligned}
[+\chi]\, p \quad &\leftrightarrow\ p; && (2.91)\\
[+\chi]\, \neg\varphi \quad &\leftrightarrow\ \neg\, [+\chi]\, \varphi; && (2.92)\\
[+\chi]\, (\varphi \wedge \psi) \quad &\leftrightarrow\ [+\chi]\, \varphi \wedge [+\chi]\, \psi; && (2.93)\\
[+\chi]\, \downarrow x.\varphi \quad &\leftrightarrow \downarrow x.[+\chi]\varphi; && (2.94)\\
[+\chi]\, \Box^{\rightarrow\!\!\!\!\cdot}\varphi \quad &\leftrightarrow\ \Box^{\rightarrow\!\!\!\!\cdot}\, [+\chi]\, \varphi; && (2.95)\\
[\uparrow \chi]\, \Box^{\succ_x}\varphi \quad &\leftrightarrow\ \Box^{\succ_x}\bot \vee (\chi \wedge \Box^{\succ_x}\neg\chi)\\
&\qquad \vee \big(\Box^{\succ_x}[\uparrow \chi]\varphi \wedge \Box^{\preceq_x}((\chi \wedge \Box^{\succ_x}\neg\chi) \rightarrow [\uparrow \chi]\varphi)\big); && (2.96)\\
[\uparrow \chi]\, \Box^{\preceq_x}\varphi \quad &\leftrightarrow\ \big((\Box^{\succ_x}\bot \vee (\chi \wedge \Box^{\succ_x}\neg\chi)) \wedge \Box^{x}[\uparrow \chi]\varphi\big)\\
&\qquad \vee\ \Box^{\preceq_x}((\chi \wedge \Box^{\succ_x}\neg\chi) \vee [\uparrow \chi]\varphi); && (2.97)\\
[\Uparrow \chi]\, \Box^{\succ_x}\varphi \quad &\leftrightarrow\ \Diamond^{\preceq_x}(\chi \wedge \Box^{\succ_x}\neg\chi)\\
&\qquad \vee \big(\neg\Diamond^{\preceq_x}(\chi \wedge \Box^{\succ_x}\neg\chi) \wedge \Box^{\succ_x}\, [\Uparrow \chi]\, \varphi\big); && (2.98)\\
[\Uparrow \chi]\, \Box^{\preceq_x}\varphi \quad &\leftrightarrow\ \big(\Diamond^{\preceq_x}(\chi \wedge \Box^{\succ_x}\neg\chi) \wedge \Box^{x}[\Uparrow \chi]\varphi\big)\\
&\qquad \vee \big(\neg\Diamond^{\preceq_x}(\chi \wedge \Box^{\succ_x}\neg\chi) \wedge \Box^{\preceq_x}\, [\Uparrow \chi]\, \varphi\big). && (2.99)
\end{aligned}
$$

Table 2.3: reduction axioms for context change

reverse the direction of the implications. It is also immediate from the truth definitions that (2.93) is valid. For (2.94) and (2.95), $[+\chi]$ and $\downarrow x.$ commute and $[+\chi]$ and $\Box^{\rightarrow\!\!\!\!\cdot}$ commute because the $+\chi$ operations do not change the assignment function g or the relation $\rightarrow\!\!\!\!\cdot$ from the initial model $\mathcal{M}$ to the new model $\mathcal{M}^{+\chi}$.

For (2.96), the lhs expresses that after context change by $\uparrow \chi$, all worlds that are more relevant than the current world w according to $\preceq^{\uparrow\chi}_{g(x)}$ satisfy φ:

$$\{v \in W \mid v \prec^{\uparrow\chi}_{g(x)} w\} \subseteq [\![\varphi]\!]^{\mathcal{M}^{\uparrow\chi}}. \tag{2.100}$$

Case 1: $\{v \in W \mid v \prec^{\uparrow\chi}_{g(x)} w\} = \emptyset$. This implies (2.100) and is equivalent to

$$w \in \mathrm{Min}_{\preceq^{\uparrow\chi}_{g(x)}}(W). \tag{2.101}$$

By Definition 2.15 for $\uparrow$, (2.101) holds iff either

$$w \in \mathrm{Min}_{\preceq_{g(x)}}(W), \tag{2.102}$$

which is equivalent to $\mathcal{M}, g, w \vDash \Box^{\succ_x} \bot$, or else

$$w \in \mathrm{Min}_{\preceq_{g(x)}}([\![\chi]\!]^{\mathcal{M}}), \tag{2.103}$$

which is equivalent to $\mathcal{M}, g, w \vDash \chi \wedge \Box^{\succ_x} \neg\chi$. This accounts for the first two disjuncts on the rhs of (2.96).

Case 2: $\{v \in W \mid v \prec^{\uparrow\chi}_{g(x)} w\} \neq \emptyset$. In this case, by Definition 2.15 for $\uparrow$,

$$\{v \in W \mid v \prec^{\uparrow\chi}_{g(x)} w\} = \{v \in W \mid v \prec_{g(x)} w\} \cup \mathrm{Min}_{\preceq_{g(x)}}([\![\chi]\!]^{\mathcal{M}}). \tag{2.104}$$

Hence (2.100) requires that

$$\{v \in W \mid v \prec_{g(x)} w\} \subseteq [\![\varphi]\!]^{\mathcal{M}^{\uparrow\chi}} = [\![[\uparrow \chi]\varphi]\!]^{\mathcal{M}}, \tag{2.105}$$

which is equivalent to $\mathcal{M}, g, w \vDash \Box^{\succ_x}[\uparrow \chi]\varphi$, and

$$\mathrm{Min}_{\preceq_{g(x)}}([\![\chi]\!]^{\mathcal{M}}) \subseteq [\![\varphi]\!]^{\mathcal{M}^{\uparrow\chi}} = [\![[\uparrow \chi]\varphi]\!]^{\mathcal{M}}, \tag{2.106}$$

which is equivalent to $\mathcal{M}, g, w \vDash \Box^{x}((\chi \wedge \Box^{\succ_x} \neg\chi) \rightarrow [\uparrow \chi]\varphi)$. The conjunction of $\Box^{\succ_x}[\uparrow \chi]\varphi$ and $\Box^{x}((\chi \wedge \Box^{\succ_x} \neg\chi) \rightarrow [\uparrow \chi]\varphi)$ is equivalent to

$$\Box^{\succ_x}[\uparrow \chi]\varphi \wedge \Box^{\preceq_x}((\chi \wedge \Box^{\succ_x} \neg\chi) \rightarrow [\uparrow \chi]\varphi), \tag{2.107}$$

which is the last disjunct on the rhs of (2.96).

For (2.97), what the lhs expresses about the current world w is

$$\{v \in W \mid w \preceq^{\uparrow\chi}_{g(x)} v\} \subseteq [\![\varphi]\!]^{\mathcal{M}^{\uparrow\chi}}. \tag{2.108}$$

Case 1: $\{v \in W \mid w \preceq^{\uparrow\chi}_{g(x)} v\} = W_{g(x)}$. This is equivalent to (2.101), which

explains the first conjunct of the first disjunct on the rhs of (2.97). In this case, (2.108) requires that

$$W_{g(x)} \subseteq [\![\varphi]\!]^{\mathcal{M}^{\uparrow\chi}} = [\![[\uparrow \chi]\varphi]\!]^{\mathcal{M}}, \tag{2.109}$$

which is equivalent to $\mathcal{M}, g, w \vDash \Box^x[\uparrow \chi]\varphi$. This accounts for the second conjunct of the first disjunct on the rhs of (2.97).

Case 2: $\{v \in W \mid w \preceq^{\uparrow\chi}_{g(x)} v\} \neq W_{g(x)}$. In this case, by Definition 2.15 for $\uparrow$,

$$\{v \in W \mid w \preceq^{\uparrow\chi}_{g(x)} v\} = \{v \in W \mid w \preceq_{g(x)} v\} \setminus \mathrm{Min}_{\preceq_{g(x)}}([\![\chi]\!]^{\mathcal{M}}). \tag{2.110}$$

Hence (2.108) requires that

$$\{v \in W \mid w \preceq_{g(x)} v\} \setminus \mathrm{Min}_{\preceq_{g(x)}}([\![\chi]\!]^{\mathcal{M}}) \subseteq [\![\varphi]\!]^{\mathcal{M}^{\uparrow\chi}} = [\![[\uparrow \chi]\varphi]\!]^{\mathcal{M}}, \tag{2.111}$$

which is equivalent to $\mathcal{M}, g, w \vDash \Box^{\preceq_x}((\chi \wedge \Box^{\succ_x}\neg\chi) \vee [\uparrow \chi]\varphi)$. This explains the second disjunct on the rhs of (2.97). The arguments for (2.98) - (2.99) are similar. □

Given Propositions 2.10 and 2.11, if we combine the epistemic and RA languages and interpret $K\varphi$ according to D-semantics (a similar point holds for L), then we can write a reduction axiom for context change and knowledge as follows:

$$[+\chi]K\psi \leftrightarrow \downarrow x.\Box^{\rightarrow}(\neg[+\chi]\sigma_d(\psi) \rightarrow \neg\alpha), \tag{2.112}$$

where α is the rhs of (2.96) if $+$ is $\uparrow$ (resp. of (2.98) if $+$ is $\Uparrow$) with $\varphi := \sigma_d(\psi)$. Here we have used the fact that $\Diamond^{\succ_x}\neg\sigma_d(\psi)$ is equivalent to $\neg\Box^{\succ_x}\sigma_d(\psi)$, and $[+\chi]\neg\Box^{\succ_x}\sigma_d(\psi)$ reduces to $\neg[+\chi]\Box^{\succ_x}\sigma_d(\psi)$, which in turn reduces to $\neg\alpha$.

An important technical and conceptual issue raised by a result like Proposition 2.11 concerns the distinction between *valid* and *schematically valid* principles of context change. Where a principle is schematically valid just in case all of its substitution instances are valid [Benthem, 2011, §3.12], the valid reduction principle $[+\chi]\,p \leftrightarrow p$ is clearly not schematically valid. Observe that $[+\chi]\,Kp \leftrightarrow Kp$ is not valid; if it were, there would be no epistemic dynamics. A more interesting example is the valid

principle $\neg Kp \to [+\chi]\neg Kp$, which holds for our operations that make the context more epistemically "demanding." Observe that $\neg K\psi \to [+\chi]\neg K\psi$ is not valid for all ψ; it is possible to count as having some knowledge after the context becomes more demanding that one did not count as having before. How can this be? The answer is that this new knowledge may be knowledge of *ignorance*.[74] This can be seen by substituting $\neg Kp$ for ψ and either trying out model changes or using (2.112) to reduce $\neg K\neg Kp \to [+\neg p]\neg K\neg Kp$ to a static principle that can be seen to be invalid. These observations raise the question, which we leave open, of what is the complete set of schematically valid principles of context change.

We leave as another open problem the task of finding an axiomatization of the theory of RA models in the static RA language (or some static extension thereof). Together with the reduction axioms of Proposition 2.11, that would give an axiomatization of the theory of RA models in the dynamic RA language to go alongside the axiomatization in the epistemic language given by Theorem 2.4.

[74]This is easiest to understand in a multi-agent setting. (Note that all of our definitions easily generalize to the multi-agent case where the modal operators in our language and relations in our models are indexed for different agents.) Taking $\psi := \neg K_j p$, suppose agent i believes of agent j that $\neg K_j p$, but i does not know $\neg K_j p$, as i has not eliminated some relevant $K_j p$-worlds. If the context changes in such a way that j no longer counts as knowing p under any circumstances, then relative to this new context, i can count as knowing $\neg K_j p$. We can no longer fault i for not having eliminated some relevant $K_j p$-worlds if there are none relative to the current context.

3

Fallibilism 1.0

In Chapter 2, I proposed formalizations of several RA and subjunctivist theories of knowledge. In this chapter, I propose a unifying framework into which all of these theories fit as special cases. The basic idea is that the RA and subjunctivist theories are all versions of what I call *ruling out fallibilism* (RO fallibilism), the view that knowing a proposition does not always require "ruling out" every last possibility in which it is false. Although subjunctivists tend not to use the RA theorists' talk of "ruling out," I will show how we can see both of their approaches as versions of RO fallibilism. Doing so involves moving from the *world-ordering* pictures of Chapter 2 to a more general *set-selection function* picture, which brings in interesting connections with the study of preference orderings and choice functions in economics.

The RA and subjunctivist theories of the last forty years occupy only a small space of the landscape of RO fallibilist theories, a space I call *Fallibilism 1.0*. After exploring this space in this chapter, I will argue in Chapter 4 that any way of arriving at a particular theory of knowledge in this space leads to one of three serious problems:

- The Problem of Vacuous Knowledge;
- The Problem of Containment;
- The Problem of Knowledge Inflation.

In Chapter 5, I will argue that we can solve all of these problems by moving to a new space of theories that I call *Fallibilism 2.0*.

3.1 Standard Alternatives Models

The starting point of Fallibilism 1.0 is Dretske's [1981] idea that for each proposition to be known, there is "a set of situations each member of which contrasts with what is [to be] known...and must be evidentially excluded if one is to know" (373). Dretske proposes that we "call the set of possible alternatives that a person must be in an evidential position to exclude (when he knows that P) the *Relevancy Set*" (371). Similarly, let us call the set of alternatives for P that the agent in question has *not* excluded the *Uneliminated Set*. As in Chapter 2, we define two *set-selection functions*:

> $\mathsf{r}_{\mathcal{C}}(P, w)$ = the set of ("relevant") possibilities that the agent *must eliminate* in order to count as knowing proposition P in world w relative to context $\mathcal{C}$;
>
> $\mathsf{u}_{\mathcal{C}}(P, w)$ = the set of ("uneliminated") possibilities that the agent has *not eliminated* as alternatives for P in world w relative to context $\mathcal{C}$.

Recall the reason for relativizing these sets to a world and a context. First, since objective features of an agent's situation in world w may affect what alternatives are relevant in w and therefore what it takes to know P in w (see Dretske 1981, 377 and DeRose 2009, 30f on "subject factors"), we allow that $\mathsf{r}(P, w)$ may differ from $\mathsf{r}(P, v)$ for a distinct world v in which the agent's situation is different. Second, if we allow—unlike Dretske—that the conversational context $\mathcal{C}$ of those attributing knowledge to the agent (or the context of assessment of a knowledge attribution, in the sense of MacFarlane 2005) can also affect what alternatives are relevant in a given world w and therefore what it takes to count as knowing P in w relative to $\mathcal{C}$ (see DeRose 2009, 30f on "attributor factors"), then we should allow that $\mathsf{r}_{\mathcal{C}}(P, w)$ may differ from $\mathsf{r}_{\mathcal{C}'}(P, w)$ for a distinct context $\mathcal{C}'$. Similarly, if one allows that what counts as eliminating an alternative may vary with context (see DeRose 2009, 30n29) or depend on the agent's situation, then our u function should depend on the context and world as well.

According to Fallibilism 1.0,[1] the agent knows P in w relative to context $\mathcal{C}$ if and only if the Relevancy Set and Uneliminated Set do not overlap:

$$\mathsf{r}_{\mathcal{C}}(P,w) \cap \mathsf{u}_{\mathcal{C}}(P,w) = \emptyset.$$

(what must be eliminated and what is uneliminated don't overlap)

Fig. 3.1 gives graphical representations of the knowledge condition violated (left) and satisfied (right). Each circle represents the entire set W of possibilities under consideration, which contains the actual world w, and the blue region represents the subset of possibilities in which P is true. The Relevancy Set and Uneliminated Set for P in context $\mathcal{C}$ are shown in red and orange, respectively, in the white *not-P* zone. If these sets overlap, as on the left, then the agent does not know P in w relative to $\mathcal{C}$; if they do not overlap, as on the right, then the agent knows P in w relative to $\mathcal{C}$.

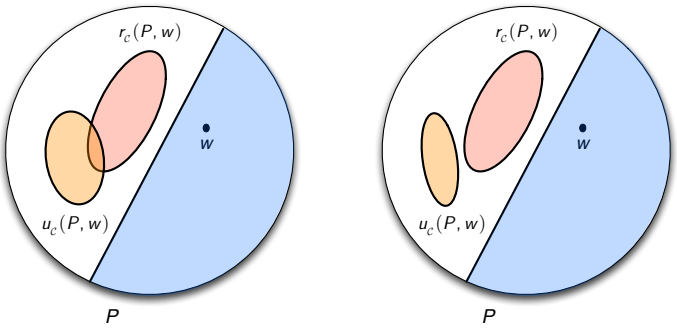

Figure 3.1: knowledge condition violated (left) vs. satisfied (right)

The pictures in Fig. 3.1 are *RO fallibilist* pictures, in the sense that the red Relevancy Set for P does not cover the entire not-P zone; according to these pictures, in order to know P, the agent need not eliminate every not-P possibility.

Claim 3.1. Standard fallibilist views fit into this set-selection function framework of Fallibilism 1.0 as special cases distinguished by the following:

- different *"structural" constraints* on the r and u functions;
- different ideas about *when an alternative must be eliminated*;

[1] As in §2.4, I omit the belief condition on knowledge for simplicity, but it is easy to add. See §3.3.2.

- different ideas about *what it is to eliminate an alternative*;
- different ideas about *what an alternative is*.

So far we have assumed a (partial) answer only to the last question: we have assumed that alternatives are possibilities/scenarios/situations/states of affairs, rather than more coarse-grained objects like *propositions*. As explained in §4.A, moving to alternatives-as-proposition will not change the main conclusions of this chapter.

The following definition formally captures the picture sketched so far.

Definition 3.1 (SA Model)**.** A *standard alternatives* (SA) model is a tuple $\mathfrak{M}$ of the form $\langle W, \mathsf{u}, \mathsf{r}, V\rangle$ where $\mathsf{u}: \mathcal{P}(W) \times W \to \mathcal{P}(W)$, $\mathsf{r}: \mathcal{P}(W) \times W \to \mathcal{P}(W)$, and as usual, W is a non-empty set and $V: \mathsf{At} \to \mathcal{P}(W)$.

We think of $\mathsf{r}(P, w)$ and $\mathsf{u}(P, w)$ as explained above, omitting the subscript for the context $\mathcal{C}$. As in Chapter 2, contextualists should think of the model $\mathfrak{M}$ as associated with a fixed context of knowledge attribution (or a fixed context of assessment), so a change in context corresponds to a change in models from $\mathfrak{M}$ to some $\mathfrak{M}'$.

The following definition states the knowledge condition that $\mathsf{r}(P, w)$ and $\mathsf{u}(P, w)$ do not overlap.

Definition 3.2 (Truth in a SA Model)**.** Given a SA model $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$ with $w \in W$ and a formula φ in the epistemic language, we define $\mathfrak{M}, w \vDash \varphi$ as follows (with propositional cases as usual):

$$\mathfrak{M}, w \vDash K\varphi \quad \text{iff} \quad \mathsf{r}(\llbracket\varphi\rrbracket^{\mathfrak{M}}, w) \cap \mathsf{u}(\llbracket\varphi\rrbracket^{\mathfrak{M}}, w) = \emptyset,$$

where $\llbracket\varphi\rrbracket^{\mathfrak{M}} = \{v \in W \mid \mathfrak{M}, v \vDash \varphi\}$.

In the next section, we will consider various constraints on the r and u functions, in line with our epistemic interpretation. Without any further constraints, the only closure property of "knowledge" in SA models is closure under logical equivalence:

$$\text{RE}\ \frac{\varphi \leftrightarrow \psi}{K\varphi \leftrightarrow K\psi}.$$

Where **E** is the weakest system of modal logic extending classical propositional logic with the RE rule, we have the following result.

Proposition 3.1 (Completeness of E)**.** **E** is sound and complete for the class of all SA models.

Proof. By the proof of Lemma 3.1 in Appendix §3.B. □

3.2 Constraints on r and u

Recall Dretske's characterization of the relevancy set for a proposition P as "a set of situations each member of which contrasts with what is [to be] known," i.e., a set of not-P situations. The following definition captures this constraint on r.

Definition 3.3 (contrast)**.** Given $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$, r satisfies contrast iff for all $w \in W$ and $P \subseteq W$,

$$\mathsf{r}(P, w) \subseteq \overline{P}\,.$$

An immediate consequence of the contrast condition is *validity omniscience*: if φ is logically valid, then φ is known.[2] For if φ is logically valid, then for any model $\mathfrak{M}$, we have $\overline{[\![\varphi]\!]}^{\mathfrak{M}} = \emptyset$, in which case $\mathsf{r}([\![\varphi]\!]^{\mathfrak{M}}, w) = \emptyset$ by the contrast condition, so $\mathsf{r}([\![\varphi]\!]^{\mathfrak{M}}, w) \cap \mathsf{u}([\![\varphi]\!]^{\mathfrak{M}}, w) = \emptyset$ no matter the value of u. Where **EN** is the weakest system of modal logic extending **E** with the necessitation rule

$$\text{N}\ \frac{\varphi}{K\varphi}\,,$$

we have the following result.

Proposition 3.2 (Completeness of EN)**.** **EN** is sound and complete for the class of SA models in which r satisfies contrast.

Proof. By the proof of Lemma 3.1 in Appendix §3.B. □

[2]Note that if we were to allow in our models "logically impossible worlds" that falsify classical validities, then the contrast condition would not imply such (classical) validity omniscience.

We have not yet imposed sufficient constraints on SA models to obtain models for *epistemic* logic, since SA models satisfying contrast do not even validate the T axiom $K\varphi \to \varphi$. One way to ensure veridicality would be to add the truth of φ as a necessary condition for the truth of $K\varphi$ in Definition 3.2. However, if we assume Lewis's [1996] *Rule of Actuality*, that "actuality is always a relevant alternative" (554), and assume that an agent can never eliminate her actual world, then as Lewis observed, veridicality follows. The following definition makes the two assumptions precise.

Definition 3.4 (Rule of Actuality)**.** Given $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$,

1. r satisfies the Rule of Actuality iff for all $w \in W$ and $P \subseteq W$,

$$\mathsf{r}\text{-RofA} \quad \text{if } w \in \overline{P}, \text{ then } w \in \mathsf{r}(P, w);$$

2. u satisfies the Rule of Actuality iff for all $w \in W$ and $P \subseteq W$,

$$\mathsf{u}\text{-RofA} \quad \text{if } w \in \overline{P}, \text{ then } w \in \mathsf{u}(P, w).$$

Given these constraints, the T axiom $K\varphi \to \varphi$ is valid. For if $w \notin [\![\varphi]\!]^{\mathfrak{M}}$, then $w \in \mathsf{r}([\![\varphi]\!]^{\mathfrak{M}}, w) \cap \mathsf{u}([\![\varphi]\!]^{\mathfrak{M}}, w)$ by the constraints, so $w \notin [\![K\varphi]\!]^{\mathfrak{M}}$. Where **ENT** is the weakest system of modal logic extending **EN** with T, we have the following result.

Proposition 3.3 (Completeness of ENT)**. ENT** is sound and complete for the class of SA models satisfying contrast, r-RofA, and u-RofA.

Proof. By the proof of Lemma 3.1 in Appendix §3.B. □

Having obtained models for an epistemic logic, let us now make the distinction between RO infallibilist and RO fallibilist views of knowledge. Simply put, RO infallibilism is the view that for every proposition P, knowing P requires ruling out *all* not-P possibilities, while RO fallibilism is the rejection of RO infallibilism.

Definition 3.5 (infallibilism and fallibilism)**.** Given $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$,

1. r satisfies infalliblism iff for all $w \in W$ and $P \subseteq W$,

$$\overline{P} \subseteq \mathsf{r}(P, w);$$

2. r satisfies fallibilism iff it does not satisfy infallibilism.

Here we are interested in RO fallibilist theories, so we will not assume infallibilism. With the freedom of fallibilism comes a number of choices about further structural properties of r. We will discuss one of the most important of these in the next section.

3.2.1 The RS and RO Parameters

In this section, I review the RS and RO theory parameters from Chapter 2.

Any RO fallibilist must answer the following questions. First, where w is the actual world and v is some possibility, can we say whether v is simply "relevant" in w, independently of any proposition in question; or must we instead say that v is relevant in w *as an alternative for* a particular proposition P, allowing that v may not be relevant in w as an alternative for a different proposition Q? Second, can we say whether v is simply "ruled out" in w, independently of any proposition in question; or must we instead say that v is ruled out in w *as an alternative for* a particular P, allowing that v may not be ruled out in w as an alternative for a different Q?

If the answer is 'yes' to the first disjunct of the first question, then there is a *fixed set* $\mathsf{R}(w) \subseteq W$ of "relevant" worlds, singled out independently of any proposition in question, such that for any proposition P, the worlds that one must rule out as alternatives for P in order to know P in w are exactly the not-P worlds in $\mathsf{R}(w)$. Similarly, if the answer is 'yes' to the first disjunct of the second question, then there is fixed set $\mathsf{U}(w) \subseteq W$ of "uneliminated" worlds, singled out independently of any proposition in question, such that for any proposition P, the worlds that one has not eliminated as alternative for P in w are exactly the not-P worlds in $\mathsf{U}(w)$.

The foregoing observations are the source of the following crucial definition.

Definition 3.6 (RS and RO Parameters)**.** Given $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V \rangle$,

1. r satisfies $\mathsf{RS}_{\exists\forall}$ iff for all $w \in W$,

   there is $(\exists)$ $\mathsf{R}(w) \subseteq W$ such that for all $(\forall)$ $P \subseteq W$, $\mathsf{r}(P, w) = \mathsf{R}(w) \cap \overline{P}$.

2. r satisfies $\mathsf{RS}_{\forall\exists}$ iff it does not satisfy $\mathsf{RS}_{\exists\forall}$.[3]

3. u satisfies $\mathsf{RO}_{\exists\forall}$ iff for all $w \in W$,

   there is $(\exists)$ $\mathsf{U}(w) \subseteq W$ such that for all $(\forall)$ $P \subseteq W$, $\mathsf{u}(P, w) = \mathsf{U}(w) \cap \overline{P}$.

4. u satisfies $\mathsf{RO}_{\forall\exists}$ iff it does not satisfy $\mathsf{RO}_{\exists\forall}$.

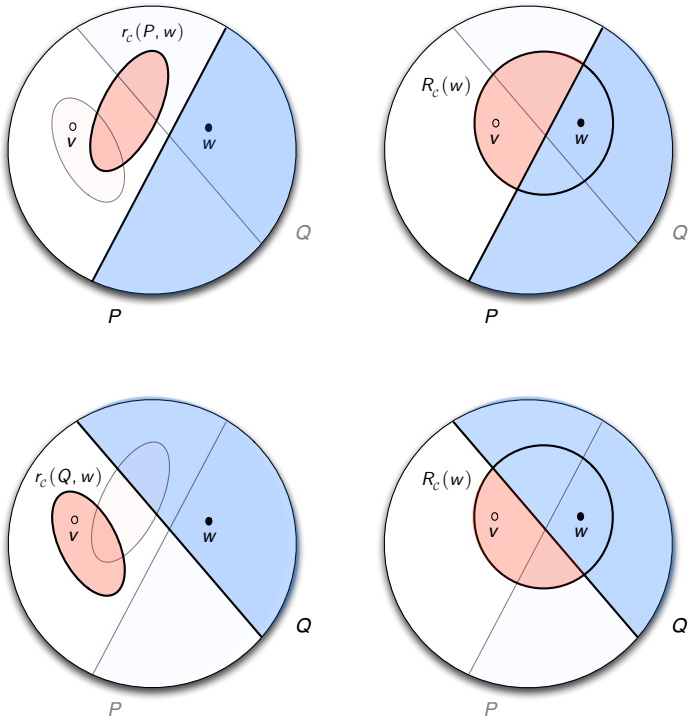

Figure 3.2: $\mathsf{RS}_{\forall\exists}$ (left) vs. $\mathsf{RS}_{\exists\forall}$ (right) parameter settings

Fig. 3.2 gives a graphical representation of the difference between $\mathsf{RS}_{\forall\exists}$ and $\mathsf{RS}_{\exists\forall}$ parameters settings. On the $\mathsf{RS}_{\forall\exists}$ side, v is a not-P world and a not-Q world, but

[3]Since $\exists\forall$ implies $\forall\exists$, our definition of $\mathsf{RS}_{\exists\forall}$ and $\mathsf{RS}_{\forall\exists}$ as mutually exclusive involves some abuse, but it makes classifications cleaner.

while v is a world that must be ruled out in order to know Q, it is not a world that must be ruled out in order to know P. By contrast, on the $\mathsf{RS}_{\exists\forall}$ side, no such split-decision on v is possible. The pictures for $\mathsf{RO}_{\forall\exists}$ vs. $\mathsf{RO}_{\exists\forall}$ would be the same if we were to substitute u for r and U for R. As observed in Chapter 2 and made precise in §3.3, the theories of Lewis [1996], Sosa [1999], DeRose [1995], Dretske [1981], Nozick [1981], and Heller [1999a] have the parameter settings in Fig. 3.3.

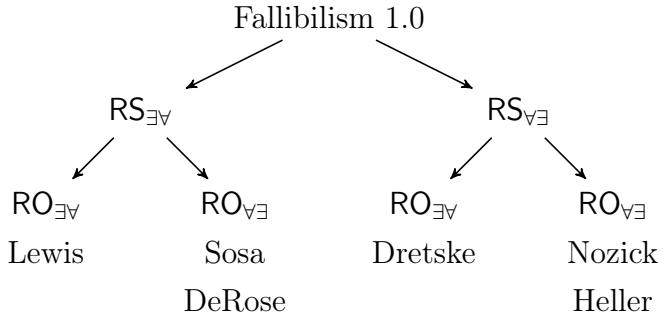

Figure 3.3: theories classified by RS and RO parameter settings

It is noteworthy that the $\mathsf{RS}_{\exists\forall}$ parameter setting is something that Lewis, Sosa, and DeRose have in common with infallibilists, as the following fact shows.

Fact 3.1 (infallibilism and $\mathsf{RS}_{\exists\forall}$)**.** If r satisfies contrast, then r satisfies infallibilism iff it satisfies $\mathsf{RS}_{\exists\forall}$ with $\mathsf{R}(w) = W$ for all $w \in W$.

As discussed in Chapters 2, the $\exists\forall$ vs. $\forall\exists$ distinctions have crucial consequences for closure. Assuming $\mathsf{RS}_{\exists\forall}$ and $\mathsf{RO}_{\exists\forall}$, the knowledge condition becomes

$$\begin{array}{ccccc} \mathsf{r}(P,w) & \cap & \mathsf{u}(P,w) & = \emptyset \\ \| & & \| & \\ \mathsf{R}(w) \cap \overline{P} & \cap & \mathsf{U}(w) \cap \overline{P} & = \emptyset, \end{array}$$

which is equivalent to

$$\mathsf{R}(w) \cap \mathsf{U}(w) \subseteq P. \tag{3.1}$$

Now it is easy to see that the K axiom $(K\varphi \wedge K(\varphi \to \psi)) \to K\psi$ is valid. For if the

agent knows φ and $\varphi \to \psi$ in w, then as instances of (3.1) we have

$$\begin{array}{lcll} \mathsf{R}(w) \cap \mathsf{U}(w) & \subseteq & [\![\varphi]\!]^{\mathfrak{M}} & \text{and} \\ \mathsf{R}(w) \cap \mathsf{U}(w) & \subseteq & [\![\varphi \to \psi]\!]^{\mathfrak{M}}. & \end{array}$$

It follows by propositional logic that

$$\mathsf{R}(w) \cap \mathsf{U}(w) \subseteq [\![\psi]\!]^{\mathfrak{M}},$$

so the agent knows ψ in w. Where **K** (resp. **KT**) is the weakest system of modal logic extending **EN** (resp. **ENT**) with the K axiom, we have the following result.

Proposition 3.4 (Completeness of **K** and **KT**)**.**

1. **K** is sound and complete for the class of SA models satisfying $\mathsf{RS}_{\exists\forall}$ and $\mathsf{RO}_{\exists\forall}$.
2. **KT** is sound and complete for the class of SA models satisfying $\mathsf{RS}_{\exists\forall}$, $\mathsf{RO}_{\exists\forall}$, r-RofA, and u-RofA.

Proof. By the proof of Lemma 3.3 in Appendix §3.B. □

If we do not assume $\mathsf{RS}_{\exists\forall}$ and $\mathsf{RO}_{\exists\forall}$, then as in Fig. 3.2, a $(\neg p \wedge \neg q)$-world v that is relevant/uneliminated as an alternative for q may not be relevant/uneliminated as an alternative for p (e.g., think of p as some mundane claim, q as the *denial* of a radical skeptical hypothesis, and v as a skeptical scenario), even if the agent knows $p \to q$, which opens up the possibility of a failure of K. As discussed in §2.1, this is one of the fundamental disagreements between Stine [1976], who insists on $\mathsf{RS}_{\exists\forall}$ and full closure, and Dretske, who allows $\mathsf{RS}_{\forall\exists}$ and some closure failure. We will return to this disagreement in §4.1, where I will raise problems for Stine's position.

3.2.2 Separating Closure Conditions

Having seen the relationship between $\mathsf{RS}_{\exists\forall}$, $\mathsf{RO}_{\exists\forall}$, and full closure under known implication—the K axiom—in §3.2.1, let us now "break apart" the conditions for full closure to obtain a more fine-grained analysis. First we will consider the condition that

corresponds, in a sense made precise in §3.2.3, to the M axiom $K(\varphi \wedge \psi) \rightarrow K\varphi \wedge K\psi$, as well as a weaker version of this condition that will play an important role in §3.3.

Definition 3.7 (cover, beta)**.** Given $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V \rangle$,

1. r satisfies cover iff for all $w \in W$ and $P, Q \subseteq W$,

$$\text{if } P \subseteq Q, \text{ then } \mathsf{r}(Q, w) \subseteq \mathsf{r}(P, w);$$

2. r satisfies beta iff for all $w \in W$ and $P, Q \subseteq W$,

$$\text{if } P \subseteq Q \text{ and } \mathsf{r}(P, w) \cap \mathsf{r}(Q, w) \neq \emptyset, \text{ then } \mathsf{r}(Q, w) \subseteq \mathsf{r}(P, w).$$

I will explain the "beta" terminology in Remark 3.1 and the significance of beta in §3.3. First, let us concentrate on the cover condition, which says that if P excludes at least as much of logical space as Q does, then coming to know P should require at least as much epistemic work, in terms of ruling out possibilities, as coming to know Q does. I will have more to say about this later, but for now let us observe that the M axiom $K(\varphi \wedge \psi) \rightarrow K\varphi \wedge K\psi$ is valid over models satisfying cover and $\mathsf{RO}_{\exists\forall}$. Intuitively, this is clear: since $\varphi \wedge \psi$ is as strong as φ, cover says that coming to know φ does not require ruling out any more possibilities than coming to know $\varphi \wedge \psi$ does. Formally, for any model $\mathfrak{M}$,

$$[\![\varphi \wedge \psi]\!]^{\mathfrak{M}} \subseteq [\![\varphi]\!]^{\mathfrak{M}}, \tag{3.2}$$

so by cover we have

$$\mathsf{r}([\![\varphi]\!]^{\mathfrak{M}}, w) \subseteq \mathsf{r}([\![\varphi \wedge \psi]\!]^{\mathfrak{M}}, w). \tag{3.3}$$

By $\mathsf{RO}_{\exists\forall}$, there is some $\mathsf{U}(w) \subseteq W$ such that:

$$\mathsf{u}([\![\varphi]\!]^{\mathfrak{M}}, w) = \mathsf{U}(w) \cap \overline{[\![\varphi]\!]}^{\mathfrak{M}}; \tag{3.4}$$

$$\mathsf{u}([\![\varphi \wedge \psi]\!]^{\mathfrak{M}}, w) = \mathsf{U}(w) \cap \overline{[\![\varphi \wedge \psi]\!]}^{\mathfrak{M}}. \tag{3.5}$$

Since $\overline{[\![\varphi]\!]}^{\mathfrak{M}} \subseteq \overline{[\![\varphi \wedge \psi]\!]}^{\mathfrak{M}}$, it follows from (3.4) - (3.5) that

$$\mathsf{u}([\![\varphi]\!]^{\mathfrak{M}}, w) \subseteq \mathsf{u}([\![\varphi \wedge \psi]\!]^{\mathfrak{M}}, w). \tag{3.6}$$

Now if $\mathfrak{M}, w \vDash K(\varphi \wedge \psi)$, then by Definition 3.2,

$$\mathsf{r}([\![\varphi \wedge \psi]\!]^{\mathfrak{M}}, w) \cap \mathsf{u}([\![\varphi \wedge \psi]\!]^{\mathfrak{M}}, w) = \emptyset, \tag{3.7}$$

which with (3.3) and (3.6) implies

$$\mathsf{r}([\![\varphi]\!]^{\mathfrak{M}}, w) \cap \mathsf{u}([\![\varphi]\!]^{\mathfrak{M}}, w) = \emptyset, \tag{3.8}$$

so $\mathfrak{M}, w \vDash K\varphi$ by Definition 3.2. The argument for $\mathfrak{M}, w \vDash K\psi$ is analogous.

I regard the validation of $K(\varphi \wedge \psi) \rightarrow K\varphi \wedge K\psi$ as a desideratum for any good theory of knowledge, so the connection between this principle and cover is noteworthy. We will return to this in §4.2 when we discuss the Problem of Containment.

Putting together our observations so far, we have the following result.

Proposition 3.5 (Completeness of EMNT)**. EMNT** is sound and complete for the class of SA models satisfying cover, $\mathsf{RO}_{\exists\forall}$, contrast, r-RofA, and u-RofA.

Proof. By the proof of Lemma 3.1 in Appendix §3.B. □

Moving on from the M axiom, let us now consider the condition that corresponds, in a sense made precise in §3.2.3, to the C axiom $K\varphi \wedge K\psi \rightarrow K(\varphi \wedge \psi)$.

Definition 3.8 (alpha)**.** Given $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V \rangle$, r satisfies alpha iff for all $w \in W$ and $P, Q \subseteq W$,

$$\mathsf{r}(P \cap Q, w) \subseteq \mathsf{r}(P, w) \cup \mathsf{r}(Q, w).$$

The argument that the C axiom is valid over models satisfying alpha and $\mathsf{RO}_{\exists\forall}$ is straightforward and similar to the argument for M and cover above. In essence, alpha says that the set of worlds one must rule out in order to know a conjunction is a subset of the set of worlds that one must rule out in order to know both conjuncts individually, which makes the connection with $K\varphi \wedge K\psi \rightarrow K(\varphi \wedge \psi)$ clear.

Proposition 3.6 (Completeness of ECNT)**. ECNT** is sound and complete for the class of SA models satisfying alpha, $\mathsf{RO}_{\exists\forall}$, contrast, r-RofA, and u-RofA.

Proof. By the proof of Lemma 3.2 in Appendix §3.B. □

Just as we have written the alpha condition in a form that makes its connection with $K\varphi \wedge K\psi \to K(\varphi \wedge \psi)$ obvious, we can rewrite the cover condition in an equivalent way that makes its connection with $K(\varphi \wedge \psi) \to K\varphi \wedge K\psi$ obvious; and we can rewrite alpha in an equivalent way that is parallel to our original formulation of cover.

Observation 3.1 (Relation of cover and alpha)**.** The following are equivalent:

cover $\forall P, Q \subseteq W$: if $P \subseteq Q$, then $\mathsf{r}(Q, w) \subseteq \mathsf{r}(P, w)$;

$\forall P, Q \subseteq W$: $\mathsf{r}(P, w) \cap \mathsf{r}(Q, w) \subseteq \mathsf{r}(P \cap Q, w)$.

Assuming contrast, the following are equivalent:[4]

$\forall P, Q \subseteq W$: if $P \subseteq Q$, then $\mathsf{r}(P, w) \cap \overline{Q} \subseteq \mathsf{r}(Q, w)$;

alpha $\forall P, Q \subseteq W$: $\mathsf{r}(P \cap Q, w) \subseteq \mathsf{r}(P, w) \cup \mathsf{r}(Q, w)$.

The reason we have written cover in the first of the two forms is to make clear that beta is a weakening of cover, an important point to which we will return in §4.2. The first forms of cover and alpha also make it clear that, mathematically speaking, cover is just the *antitonicity* of r, while alpha is a kind of weakening of monotoniticity.

Recall that the following are interderivable using the RE rule: (i) the combination of the M axiom $K(\varphi \wedge \psi) \to K\varphi \wedge K\psi$ and the C axiom $K\varphi \wedge K\psi \to K(\varphi \wedge \psi)$ and (ii) the K axiom $(K\varphi \wedge K(\varphi \to \psi)) \to K\psi$. A parallel point applies to the corresponding conditions on r: the combination of cover and alpha is equivalent to $\mathsf{RS}_{\exists\forall}$.

[4]From the first to the second, given $P \cap Q \subseteq P$ and $P \cap Q \subseteq Q$, the first condition implies

$$\mathsf{r}(P \cap Q, w) \cap \overline{P} \subseteq \mathsf{r}(P) \text{ and } \mathsf{r}(P \cap Q, w) \cap \overline{Q} \subseteq \mathsf{r}(Q). \tag{3.9}$$

Given contrast, $\mathsf{r}(P \cap Q, w) \subseteq \overline{\overline{P} \cap \overline{Q}} = \overline{P} \cup \overline{Q}$, which with (3.9) implies $\mathsf{r}(P \cap Q, w) \subseteq \mathsf{r}(P, w) \cup \mathsf{r}(Q, w)$.
From the second to the first, given $P \subseteq Q$ and the second condition, we have

$$\mathsf{r}(P, w) = \mathsf{r}(Q \cap (P \cup \overline{Q}), w) \subseteq \mathsf{r}(Q, w) \cup \mathsf{r}(P \cup \overline{Q}, w). \tag{3.10}$$

Given contrast, $\mathsf{r}(P \cup \overline{Q}, w) \subseteq \overline{\overline{P} \cap Q} = \overline{P} \cap Q$, which with (3.10) implies $\mathsf{r}(P, w) \cap \overline{Q} \subseteq \mathsf{r}(Q, w)$.

Fact 3.2 ($\mathsf{RS}_{\exists\forall}$ = cover + alpha)**.** r satisfies $\mathsf{RS}_{\exists\forall}$ iff r satisfies cover and alpha.

Proof. The left-to-right direction is an easy exercise. For the right-to-left direction, for each $w \in W$ we must present some $\mathsf{R}(w) \subseteq W$ such that for all $P \subseteq W$,

$$\mathsf{r}(P, w) = \mathsf{R}(w) \cap \overline{P}. \tag{3.11}$$

The desired set is given by

$$\mathsf{R}(w) = \mathsf{r}(\emptyset, w). \tag{3.12}$$

Since $\emptyset \subseteq P$, we have $\mathsf{r}(P, w) \subseteq \mathsf{r}(\emptyset, w)$ by cover, so $\mathsf{r}(P, w) \subseteq \mathsf{r}(\emptyset, w) \cap \overline{P}$ given contrast. For the other inclusion, since $\emptyset \subseteq P$, we have $\mathsf{r}(\emptyset, w) \cap \overline{P} \subseteq \mathsf{r}(P, w)$ by the alternative version of alpha given in Observation 3.1. Hence $\mathsf{r}(P, w) = \mathsf{r}(\emptyset, w) \cap \overline{P}$. □

As desired, Fact 3.2 "breaks up" the conditions for full closure. We will return to the significance of this when we discuss the Problem of Containment in §4.2.

Choice Functions in Economics

Those familiar with the literature on choice functions in economics will have noticed the connection between the constraints we have considered for r and standard constraints proposed for rational choice functions. To make this precise, we need to consider one more constraint on r, which will play a fundamental role in §4.1.

Definition 3.9 (no vacuous knowledge)**.** Given $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V \rangle$, r satisfies no vacuous knowledge iff for all $w \in W$ and $P \subseteq W$,

$$\mathsf{noVK} \quad \text{if } P \neq W, \text{ then } \mathsf{r}(P, w) \neq \emptyset.$$

In other words, noVK says that if P is contingent, then knowledge of P cannot come "for free," in the sense of not requiring the elimination of any possibilities. I will say much more about this later, but for now let us observe that if r satisfies noVK and contrast, then we can define a *choice function* c based on r as follows.

Definition 3.10 (c function)**.** Given $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$, for all $P \subseteq W$ and $w \in W$,

$$\mathsf{c}(P, w) = \mathsf{r}(\overline{P}, w).$$

Think of c as choosing, for any set P, the relevant P-worlds. It is what economists call a choice function just in case for any *non-empty* set P, it chooses a *non-empty* set of relevant P-worlds: if $P \neq \emptyset$, then $\emptyset \neq \mathsf{c}(P, w) \subseteq P$, which holds just in case r satisfies noVK and contrast. With this we can make the economics connection.

Remark 3.1 (Choice Functions)**.** Using Definition 3.10, the alpha condition of Definition 3.8 can be restated as $\mathsf{c}(X \cup Y, w) \subseteq \mathsf{c}(X, w) \cup \mathsf{c}(Y, w)$. This is equivalent to what is known in the economics literature on choice functions as the "Chernoff condition" [Chernoff, 1954]: if $X \subseteq Y$, then $X \cap \mathsf{c}(Y, w) \subseteq \mathsf{c}(X, w)$. Similarly, the beta condition of Definition 3.7.2 can be restated with c as follows: if $X \subseteq Y$ and $\mathsf{c}(X, w) \cap c(Y, w) \neq \emptyset$, then $\mathsf{c}(X, w) \subseteq \mathsf{r}(Y, w)$. The alpha and beta ($\alpha$ and β) terminology is due to Sen [1971], although we have written the alpha condition in the equivalent form that Sen [1971, §9, n1] calls α^*, and we have written the beta condition in the form given by Bordes [1976, §2]. Sen observed that the conjunction of his α and β is equivalent to the well-known "Arrow condition" [Arrow, 1959, §2]:

$$\text{if } X \subseteq Y \text{ and } X \cap \mathsf{c}(Y, w) \neq \emptyset, \text{ then } \mathsf{c}(X, w) = X \cap \mathsf{c}(Y, w)\,.$$

Similarly, we could repackage our alpha and beta into a single condition Arrow:

$$\text{if } X \subseteq Y \text{ and } \mathsf{r}(X, w) \cap \overline{Y} \neq \emptyset, \text{ then } \mathsf{r}(Y, w) = \mathsf{r}(X, w) \cap \overline{Y}.$$

Finally, the strengthening of beta called cover is equivalent to Sen's γ^* [1971, §9, n1]. For further discussion of choice functions and their relation to orderings, to which we will return in §3.3.1, see Rott 2001, Ch. 6.

Return of the X Axiom

As noted above and made precise in §3.2.3, alpha corresponds to the C axiom $K\varphi \wedge K\psi \rightarrow K(\varphi \wedge \psi)$, while the strengthening of beta in cover corresponds to the M axiom $K(\varphi \wedge \psi) \rightarrow K\varphi \wedge K\psi$. What about the other key axiom from Chapter 2, the X axiom $K(\varphi \wedge \psi) \rightarrow K\varphi \vee K\psi$? The corresponding condition appears in the literature on counterfactuals as condition (e) in Loewer 1979 and Mayer 1981.

Definition 3.11 ((e) condition)**.** Given $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$, r satisfies the (e) condition iff for all $w \in W$ and $P \subseteq W$,

$$\text{(e)} \quad \mathsf{r}(P, w) \subseteq \mathsf{r}(P \cap Q, w) \text{ or } \mathsf{r}(Q, w) \subseteq \mathsf{r}(P \cap Q, w).$$

The argument that the X axiom is valid over models satisfying (e) and $\mathsf{RO}_{\exists\forall}$ is straightforward. In essence, (e) says that for any conjunction, at least one of the conjuncts is such that the set of worlds that one must rule out in order to know that conjunct is a subset of the set of worlds that one must rule out in order to know the conjunction, which makes the connection with $K(\varphi \wedge \psi) \rightarrow K\varphi \vee K\psi$ clear.

Recall the beta condition from Definition 3.7:

$$\text{if } X \subseteq Y \text{ and } \mathsf{r}(X, w) \cap \mathsf{r}(Y, w) \neq \emptyset, \text{ then } \mathsf{r}(Y, w) \subseteq \mathsf{r}(X, w).$$

Although by itself beta—unlike alpha—does not correspond to any closure principle statable in our epistemic language, together alpha and beta imply the (e) condition that corresponds to the X axiom, which hints at the role that beta will play later.

Observation 3.2 (Relations of (e) and alpha + beta)**.**

1. alpha and beta jointly (but do not individually) imply (e).
2. alpha and (e) do not jointly imply beta.
3. beta and (e) do not jointly imply alpha.

Proof. Given $P \cap Q \subseteq P$ and $P \cap Q \subseteq Q$, by beta we have:

$$\text{if } \mathsf{r}(P \cap Q, w) \cap \mathsf{r}(P, w) \neq \emptyset, \text{ then } \mathsf{r}(P, w) \subseteq \mathsf{r}(P \cap Q, w); \tag{3.13}$$

$$\text{if } \mathsf{r}(P \cap Q, w) \cap \mathsf{r}(Q, w) \neq \emptyset, \text{ then } \mathsf{r}(Q, w) \subseteq \mathsf{r}(P \cap Q, w). \tag{3.14}$$

By alpha we have

$$\mathsf{r}(P \cap Q, w) \subseteq \mathsf{r}(P, w) \cup \mathsf{r}(Q, w), \tag{3.15}$$

and together (3.13) - (3.15) imply (e). We leave the other parts to the reader. □

We have now introduced all of the constraints on r to be considered in this chapter, and we are almost ready to see how the theories of Chapter 2 fit into this framework. Before doing so, however, let us make precise our talk of "correspondence."

3.2.3 Correspondence Theory

In the previous sections, I suggested that certain properties of r and u "correspond" to certain closure principles. We can state this formally by extending standard notions from modal correspondence theory [van Benthem, 2001] to our SA framework.

Definition 3.12 (Frames and Validity)**.** An r-*frame* is a pair $\langle W, \mathsf{r}\rangle$ where W and r are as in Definition 3.1. A formula φ is *valid on the frame* $\langle W, \mathsf{r}\rangle$ iff for all models $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$ based on the frame and all $w \in W$, $\mathfrak{M}, w \vDash \varphi$. A formula φ is valid on the frame $\langle W, \mathsf{r}\rangle$ *relative to models in class* K iff for all $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle \in \mathsf{K}$ and all $w \in W$, $\mathfrak{M}, w \vDash \varphi$. The definitions for u-*frames* are analogous.

Intuitively, a r-frame (relevance frame) is simply a set of worlds together with information about which possibilities are relevant, provided by r, but *without* information about which possibilities are (un)eliminated, provided by u, or about which atoms $p, q, r, \ldots$ hold at different worlds, provided by V. A formula φ is valid on the frame just in case no matter what information we add to the frame about which possibilities are (un)eliminated or which atoms hold where, φ will be true everywhere.

The following proposition supports our earlier correspondence claims.

Proposition 3.7 (Correspondence)**.** Let $\langle W, \mathsf{r}\rangle$ be a r-frame satisfying contrast.

1. $K\varphi \to \varphi$ is valid on $\langle W, \mathsf{r}\rangle$ relative to models satisfying u-RofA iff r satisfies r-RofA.

2. $K\varphi \wedge K\psi \to K(\varphi \wedge \psi)$ is valid on $\langle W, \mathsf{r}\rangle$ relative to models satisfying $\mathsf{RO}_{\exists\forall}$ iff r satisfies alpha.

3. $K(\varphi \wedge \psi) \to K\varphi \wedge K\psi$ is valid on $\langle W, \mathsf{r}\rangle$ relative to models satisfying $\mathsf{RO}_{\exists\forall}$ iff r satisfies cover.

4. $K(\varphi \wedge \psi) \to K\varphi \vee K\psi$ is valid on $\langle W, \mathsf{r}\rangle$ relative to models satisfying $\mathsf{RO}_{\exists\forall}$ iff r satisfies (e).

Proof. The proofs apply the standard style of modal correspondence reasoning (see, e.g., van Benthem 2010, §9.2) to our alternatives frames. We have already seen the right-to-left direction of 3, and we may prove the left-to-right by contraposition: assuming $\langle W, \mathsf{r}\rangle$ does not satisfy cover, we build a model $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$ satisfying $\mathsf{RO}_{\exists\forall}$ such that for some $w \in W$, $\mathfrak{M}, w \nvDash K(\varphi \wedge \psi) \to K\varphi \wedge K\psi$. By the assumption that r does not satisfy cover, there are $P \subseteq Q \subseteq W$ such that for some $w, v \in W$, both $v \notin \mathsf{r}(P, w)$ and $v \in \mathsf{r}(Q, w)$. First, define u such that for all $S \subseteq W$ and $x \in W$,

$$\mathsf{u}(S, x) = \{v\} \cap \overline{S}, \tag{3.16}$$

so u satisfies $\mathsf{RO}_{\exists\forall}$.[5] Since $v \notin \mathsf{r}(P, w)$, (3.16) implies

$$\mathsf{r}(P, w) \cap \mathsf{u}(P, w) = \emptyset; \tag{3.17}$$

and since $v \in \mathsf{r}(Q, w) \subseteq \overline{Q}$ given contrast, (3.16) implies

$$v \in \mathsf{r}(Q, w) \cap \mathsf{u}(Q, w) \neq \emptyset. \tag{3.18}$$

[5]So defined, u does not satisfy u-RofA. Relative to models in which u satisfies both $\mathsf{RO}_{\exists\forall}$ and u-RofA, $K(\varphi \wedge \psi) \to (K\varphi \wedge K\psi)$ is valid on $\langle W, \mathsf{r}\rangle$ iff r satisfies cover for knowable propositions, where this is the cover condition restricted to P such that $w \notin \mathsf{r}(P, w)$. With this modification, the assumption of our proof is that there are some $P \subseteq Q \subseteq W$ such that $w \notin \mathsf{r}(P, w)$, $v \in \mathsf{r}(Q, w)$, and $v \notin \mathsf{r}(P, w)$. We can then define $\mathsf{u}(S, x) = \{w, v\} \cap \overline{S}$ and (3.17) - (3.18) will hold. However, if we only assume that r does not satisfy cover, then we cannot always construct a model $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$ satisfying $\mathsf{RO}_{\exists\forall}$ and u-RofA such that $\mathfrak{M}, w \nvDash K(\varphi \wedge \psi) \to K\varphi \wedge K\psi$ for some $w \in W$.

Second, define the valuation V such that for some atomic sentences p and q, $V(p) = P$ and $V(q) = Q$. Then given $P \subseteq Q$, we have $[\![p]\!]^{\mathfrak{M}} \subseteq [\![q]\!]^{\mathfrak{M}}$, in which case $[\![p \wedge q]\!]^{\mathfrak{M}} = [\![p]\!]^{\mathfrak{M}} \cap [\![q]\!]^{\mathfrak{M}} = [\![p]\!]^{\mathfrak{M}} = P$. Together with (3.17) and (3.18) this implies

$$\begin{aligned} \mathsf{r}([\![p \wedge q]\!]^{\mathfrak{M}}, w) \cap \mathsf{u}([\![p \wedge q]\!]^{\mathfrak{M}}, w) &= \emptyset \text{ and} \\ \mathsf{r}([\![q]\!]^{\mathfrak{M}}, w) \cap \mathsf{u}([\![q]\!]^{\mathfrak{M}}, w) &\neq \emptyset, \end{aligned}$$

so $\mathfrak{M}, w \vDash K(p \wedge q)$ and $\mathfrak{M}, w \nvDash Kq$. The reasoning for the other parts is similar. □

In §4.2, we will return to these facts in connection with the Problem of Containment.

As a logical observation, if we put Propositions 3.7.2 and 3.7.4 together with Fact 3.2.2, we can see that the C and X axioms, which form the core of the Logic of Ranked Relevant Alternatives in §2.8, are not enough to "force" the r function to satisfy beta. As we will see in §3.3.1, beta is required for r to be "equivalent" to a *ranking* of worlds. What this means is that the C and X axioms do not force r to be equivalent to a ranking. As a speculation, this may explain why defining a standard canonical model "all at once" for the Logic of Ranked Relevant Alternatives is difficult: one would have to force the canonical relation $\preceq^c_w$ to be a ranking "by hand," as the axioms will not do it by themselves. By building up falsifying models inductively as in §2.6.2, we can more easily ensure that the relation is a ranking along the way.

3.3 Unification

We are now ready to establish one of the main claims made at the beginning of this chapter: all of the RA and subjunctivist theories formalized in the world-ordering pictures of Chapter 2 fit into the set-selection function picture of this chapter as special cases. In other words, these theories really do belong on the tree of Fallibilism 1.0 in Fig. 3.3. This unification will be essential for the argument of Chapter 4.

3.3.1 Relevant Alternatives

Recall that in §2.4, we formalized three different RA theories of knowledge over a single class of RA models of the form $\mathcal{M} = \langle W, \rightarrow, \preceq, V\rangle$ by defining *three different truth clauses* for knowledge formulas of the form $K\varphi$, given by C-semantics (for **C**artesian), D-semantics (for **D**retske), and L-semantics (for **L**ewis):

$$\begin{array}{lll} \mathcal{M}, w \vDash_c K\varphi & \text{iff} & \overline{[\![\varphi]\!]}_c \cap {\rightarrow}(w) = \emptyset; \\ \mathcal{M}, w \vDash_d K\varphi & \text{iff} & \mathrm{Min}_{\preceq_w}\big(\overline{[\![\varphi]\!]}_d\big) \cap {\rightarrow}(w) = \emptyset; \\ \mathcal{M}, w \vDash_l K\varphi & \text{iff} & \mathrm{Min}_{\preceq_w}(W) \cap \overline{[\![\varphi]\!]}_l \cap {\rightarrow}(w) = \emptyset. \end{array}$$

I have written these truth clauses in a different but equivalent form relative to Definition 2.5. Recall the meaning of our notation (Notation 2.1 and 2.4): for any proposition $P \subseteq W$, $\mathrm{Min}_{\preceq_w}(P)$ is the set of most relevant (at w) P-worlds;[6] and ${\rightarrow}(w)$ is the set of worlds that are uneliminated by the agent in w.

Our first step toward unification is to define *three different classes of SA models* that will "capture" C-, D-, and L-semantics over RA models, respectively, while using only the single truth clause for $K\varphi$ formulas given by Definition 3.2 in this chapter. Observe that there are three natural ways of transforming a given RA model $\mathcal{M} = \langle W, \rightarrow, \preceq, V\rangle$ into an SA model $\mathfrak{M}_x = \langle W, \mathsf{u}, \mathsf{r}_x, V\rangle$, where $x \in \{c, d, l\}$:

$$\mathsf{u}(P, w) = {\rightarrow}(w) \cap \overline{P}; \tag{3.19}$$

$$\mathsf{r}_c(P, w) = \overline{P}; \tag{3.20}$$

$$\mathsf{r}_d(P, w) = \mathrm{Min}_{\preceq_w}(\overline{P}); \tag{3.21}$$

$$\mathsf{r}_l(P, w) = \mathrm{Min}_{\preceq_w}(W) \cap \overline{P}. \tag{3.22}$$

[6] For simplicity I assume in this section that $\preceq_w$ is *total*.

Now if we apply the truth clause for $K\varphi$ used in this chapter, we have

$$\begin{array}{lll} \mathfrak{M}_c, w \vDash K\varphi & \text{iff} & \begin{array}{ccccc} \mathsf{r}_c([\![\varphi]\!], w) & \cap & \mathsf{u}([\![\varphi]\!], w) & = & \emptyset \\ \| & & \| & & \\ \overline{[\![\varphi]\!]} & \cap & \rightarrowtriangle(w) \cap \overline{[\![\varphi]\!]} & = & \emptyset; \end{array} \\ \\ \mathfrak{M}_d, w \vDash K\varphi & \text{iff} & \begin{array}{ccccc} \mathsf{r}_d([\![\varphi]\!], w) & \cap & \mathsf{u}([\![\varphi]\!], w) & = & \emptyset \\ \| & & \| & & \\ \mathrm{Min}_{\preceq_w}(\overline{[\![\varphi]\!]}) & \cap & \rightarrowtriangle(w) \cap \overline{[\![\varphi]\!]} & = & \emptyset; \end{array} \\ \\ \mathfrak{M}_l, w \vDash K\varphi & \text{iff} & \begin{array}{ccccc} \mathsf{r}_l([\![\varphi]\!], w) & \cap & \mathsf{u}([\![\varphi]\!], w) & = & \emptyset \\ \| & & \| & & \\ \mathrm{Min}_{\preceq_w}(W) \cap \overline{[\![\varphi]\!]} & \cap & \rightarrowtriangle(w) \cap \overline{[\![\varphi]\!]} & = & \emptyset. \end{array} \end{array}$$

Observe that the second, fourth, and sixth equations are equivalent to the right-hand sides of the C-, D-, and L-semantics clauses for $K\varphi$. What this shows is that if we transform a RA model $\mathcal{M}$ into a SA model $\mathfrak{M}_x$ using (3.19) - (3.22), then what the agent knows in $\mathcal{M}$ according to X-semantics will exactly agree with what the agent knows in $\mathfrak{M}_x$ according to the semantics of this chapter. The two models are "epistemically equivalent." The question now becomes: what kind of conditions on r and u does $\mathfrak{M}_x$ satisfy? In particular, we would like conditions such that if a SA model satisfies these conditions, then we can transform it into an epistemically equivalent RA model. The following definition identifies precisely such conditions.

Definition 3.13 (RA-like Model Classes)**.**

1. Let C be the class of SA models satisfying the following conditions:[7]

contrast	$\mathsf{r}(P, w) \subseteq \overline{P}$;
infallibilism	$\overline{P} \subseteq \mathsf{r}(P, w)$;
u-RofA	if $w \notin P$, then $w \in \mathsf{u}(P, w)$;
$\mathsf{RO}_{\exists\forall}$	there is $\mathsf{U}(w) \subseteq W$ such that for all $P \subseteq W$, $\mathsf{u}(P, w) = \overline{P} \cap \mathsf{U}(w)$.

[7]From now on I leave the universal quantification over $w \in W$ and $P, Q \subseteq W$ implicit.

2. Let L be the class of SA models satisfying $\mathsf{u\text{-}RofA}$, $\mathsf{RO}_{\exists\forall}$, and the following:

   $\mathsf{r\text{-}RofA}$ if $w \notin P$, then $w \in \mathsf{r}(P, w)$;

   $\mathsf{RS}_{\exists\forall}$ there is $\mathsf{R}(w) \subseteq W$ such that for all $P \subseteq W$, $\mathsf{r}(P, w) = \overline{P} \cap \mathsf{R}(w)$.

3. Let D be the class of SA models satisfying $\mathsf{u\text{-}RofA}$, $\mathsf{RO}_{\exists\forall}$, and the following:

   $\mathsf{contrast}$ $\mathsf{r}(P, w) \subseteq \overline{P}$;

   $\mathsf{r\text{-}RofA}$ if $w \notin P$, then $w \in \mathsf{r}(P, w)$;

   noVK if $P \neq W$, then $\mathsf{r}(P, w) \neq \emptyset$;

   alpha $\mathsf{r}(P \cap Q, w) \subseteq \mathsf{r}(P, w) \cup \mathsf{r}(Q, w)$;

   beta if $P \subseteq Q$ and $\mathsf{r}(P, w) \cap \mathsf{r}(Q, w) \neq \emptyset$, then $\mathsf{r}(Q, w) \subseteq \mathsf{r}(P, w)$.

Class C reflects a simple infallibilist picture: in order to know a proposition P, one must eliminate *all* not-P worlds ($\mathsf{contrast}$ + $\mathsf{infallibilism}$); an agent in w can never eliminate her actual world w ($\mathsf{u\text{-}RofA}$); and for every world w, there is a fixed (proposition-independent) set of worlds uneliminated by the agent in w ($\mathsf{RO}_{\exists\forall}$).

Class L reflects Lewis's [1996] simple fallibilist picture (for a single context): there is a fixed set of worlds that are relevant for the agent w ($\mathsf{RS}_{\exists\forall}$) and a fixed set of worlds that are uneliminated by the agent in w ($\mathsf{RO}_{\exists\forall}$); and the actual world w is always relevant for and uneliminated by the agent in w ($\mathsf{r\text{-}RofA}$ + $\mathsf{u\text{-}RofA}$).

Class D reflects a fallibilist picture combining a Lewisian ($\mathsf{RO}_{\exists\forall}$) view of ruling out with a version of the Dretskean ($\mathsf{RS}_{\forall\exists}$) view of relevant alternatives: there is a fixed set of worlds that are uneliminated by the agent in w ($\mathsf{RO}_{\exists\forall}$), and the actual world w is always uneliminated by and relevant for the agent in w ($\mathsf{u\text{-}RofA}$ + $\mathsf{r\text{-}RofA}$); the worlds that one must eliminate in order to know P are not-P worlds ($\mathsf{contrast}$); if P is contingent, then knowing P requires eliminating some world(s) (noVK); together the relevant alternatives for P and the relevant alternatives for Q include the relevant alternatives for P *and* Q (alpha); and if P is as strong as Q, and the relevant alternatives for P and the relevant alternatives for Q overlap, then the relevant alternatives for P include the relevant alternatives for Q (beta).

As suggested above, the point of defining these model classes is to obtain results of the following form: given a SA model $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$ in class D, we can transform $\mathfrak{M}$ into an RA model $\mathcal{M} = \langle W, \rightarrow\!\!\!\!\triangleright, \preceq, V\rangle$ such that when we view $\mathcal{M}$ from the perspective

of D-semantics, $\mathfrak{M}$ and $\mathcal{M}$ are epistemically equivalent. To do so, we define:

$$x \twoheadrightarrow y \quad \text{iff} \quad y \in \mathsf{U}(x), \text{ and} \tag{3.23}$$

$$x \preceq_w y \quad \text{iff} \quad \exists Q \subseteq W\colon x \in \mathsf{r}(Q, w) \text{ and } y \in \overline{Q}, \tag{3.24}$$

where $\mathsf{U}(x)$ is the set of worlds whose existence is guaranteed by $\mathsf{RO}_{\exists\forall}$ for u. The idea of (3.24) is that if x is a not-Q world that must be ruled out in order to know Q, and y is also a not-Q world, then x must be at least as relevant as y, since only the most relevant not-Q worlds must be ruled out in order to know Q. The crucial observation, essentially due to Arrow [1959], is that if r satisfies $\mathsf{contrast}$, noVK, alpha, and beta, then $\preceq_w$ defined by (3.24) is a *ranking* of worlds and the Relevancy Set $\mathsf{r}(P, w)$ is the same as the set $\mathrm{Min}_{\preceq_w}(\overline{P})$ of most relevant $\overline{P}$-worlds according to $\preceq_w$; conversely, if $\preceq_w$ is a ranking of worlds, then the function r obtained by (3.21) satisfies $\mathsf{contrast}$, noVK, alpha, and beta. Appendix §3.A contains proofs of these claims. What they show is that $\mathsf{contrast}$, noVK, alpha, and beta are the assumption about r built into Heller's [1989, 1999a] world-ordering version of the RA theory. Indeed, we have the following connections between pairs of properties bridging SA and RA models:

- r-RofA for r and weak centering for each $\preceq_w$;
- noVK for r and well-foundedness for each $\preceq_w$;
- $\mathsf{contrast} + \mathsf{noVK} + \mathsf{alpha} + \mathsf{beta}$ for r and totality + transitivity for $\preceq_w$;
- u-RofA for u and reflexivity for $\twoheadrightarrow$.

We can now state the result that we have been building up to: Theorem 3.1 shows that the C/D/L-semantics of Chapter 2 over RA models are equivalent, as semantics for the epistemic language, to our new semantics over SA models in $\mathsf{C}/\mathsf{D}/\mathsf{L}$.[8] Therefore, the completeness theorems of Chapter 2 transfer to the SA model classes.

Theorem 3.1 (RA and SA Models)**.** Let $x \in \{c, d, l\}$ and $\mathsf{X} \in \{\mathsf{C}, \mathsf{D}, \mathsf{L}\}$.

[8]Compare the result for D-semantics to the results for counterfactuals in Lewis 1973, 58f, 49f.

1. For every total, well-founded RA model $\mathcal{M}$ with the universal field property,[9] there is a SA model $\mathfrak{M}_x \in \mathsf{X}$ such that for all epistemic formulas φ,

$$\mathcal{M}, w \vDash_x \varphi \text{ iff } \mathfrak{M}_x, w \vDash \varphi.$$

2. For every SA model $\mathfrak{M} \in \mathsf{X}$, there is a total, well-founded RA model $\mathcal{M}$ with the universal field property such that for all epistemic formulas φ,

$$\mathfrak{M}, w \vDash \varphi \text{ iff } \mathcal{M}, w \vDash_x \varphi.$$

I give the proof of Theorem 3.1 in Appendix §3.A. In addition, in Remarks 3.3 and 3.4 of Appendix §3.A, I discuss the appropriate conditions for Theorem 3.1 in the cases of well-founded RA models that are not assumed to be *total* or *universal*.

3.3.2 Counterfactuals and Beliefs

Our next goal is to show that the theories of knowledge of Heller [1999a], Nozick [1981], and Sosa [1999] also fit into the framework of this chapter as special cases. Recall that in §2.5 we formalized these theories over a single class of CB models of the form $\mathcal{M} = \langle W, D, \leqslant, V\rangle$ by defining three different truth clauses, given by H-semantics (for **H**eller), N-semantics (for **N**ozick), and S-semantics (for **S**osa):

$$\begin{aligned}
&\mathcal{M}, w \vDash_x B\varphi \quad \text{iff} \quad D(w) \subseteq [\![\varphi]\!]_x;\\
&\mathcal{M}, w \vDash_h K\varphi \quad \text{iff} \quad D(w) \subseteq [\![\varphi]\!]_h \text{ and}\\
&\qquad\qquad\quad \text{(sensitivity) } \mathrm{Min}_{\leqslant_w}\big(\overline{[\![\varphi]\!]_h}\big) \cap [\![B\varphi]\!]_h = \emptyset;\\
&\mathcal{M}, w \vDash_n K\varphi \quad \text{iff} \quad D(w) \subseteq [\![\varphi]\!]_n \text{ and}\\
&\qquad\qquad\quad \text{(sensitivity) } \mathrm{Min}_{\leqslant_w}\big(\overline{[\![\varphi]\!]_n}\big) \cap [\![B\varphi]\!]_n = \emptyset,\\
&\qquad\qquad\quad \text{(adherence) } \mathrm{Min}_{\leqslant_w}(W) \cap [\![\varphi]\!]_n \cap \overline{[\![B\varphi]\!]_n} = \emptyset;
\end{aligned}$$

[9] Recall from Definition 2.3 that the universal field property requires that the field W_w of each $\preceq_w$ be W.

$$\mathcal{M}, w \models_s K\varphi \quad \text{iff} \quad D(w) \subseteq [\![\varphi]\!]_s \text{ and } \\ \text{(safety) } \mathrm{Min}_{\leqslant_w}(W) \cap \overline{[\![\varphi]\!]}_s \cap [\![B\varphi]\!]_s = \emptyset.$$

I have written these truth clauses in a different but equivalent form relative to Definition 2.7. Recall the meaning of our notation (Notation 2.1 and 2.5): for any proposition $P \subseteq W$, $\mathrm{Min}_{\leqslant_w}(P)$ is the set of closest (to w) P-worlds;[10] and $D(w)$ is the set of doxastically accessible worlds, compatible with the agent's beliefs in w. As in Chapter 2, we take $D(w) \subseteq P$ to mean that in world w, the agent believes P.

Since belief is a necessary condition for knowledge in H/N/S-semantics, let us now add belief as a necessary condition for knowledge in the framework of this chapter as well. To do so, we first add a doxastic accessibility relation D to our SA models.

Definition 3.14 (SAB Model)**.** A *standard alternatives and belief* model is a tuple $\mathfrak{M} = \langle W, D, \mathsf{u}, \mathsf{r}, V\rangle$ where $\langle W, \mathsf{u}, \mathsf{r}, V\rangle$ is a SA model as in Definition 3.1, and D is a serial binary relation on W.

We can now add the belief requirement to the $K\varphi$ clause of Definition 3.2.

Definition 3.15 (Truth in a SAB Model)**.** Given a SAB model $\mathfrak{M} = \langle W, D, \mathsf{u}, \mathsf{r}, V\rangle$ with $w \in W$ and a formula φ in the epistemic language, we define $\mathfrak{M}, w \vDash \varphi$ as follows (with propositional cases as usual):

$$\mathfrak{M}, w \vDash K\varphi \quad \text{iff} \quad D(w) \subseteq [\![\varphi]\!]^{\mathfrak{M}} \text{ and } \mathsf{r}([\![\varphi]\!]^{\mathfrak{M}}, w) \cap \mathsf{u}([\![\varphi]\!]^{\mathfrak{M}}, w) = \emptyset.$$

Toward obtaining a result for SAB and CB models analogous to Theorem 3.1 for SA and RA models, we will first treat H- and S-semantics. Dealing with the adherence condition of N-semantics requires a further generalization provided below.

As in §3.3.1, our first step to unification is to define two classes of SA models that will "capture" H- and S-semantics over RA models, respectively, while using only the single truth clause for $K\varphi$ formulas in Definition 3.15. Observe that while the *sensitivity* condition of H/N-semantics is typically thought of as the counterfactual requirement $\neg\varphi \boxdot\!\!\rightarrow \neg B\varphi$, we can also think of it in terms of a requirement that

[10] For simplicity I assume in this section that $\leqslant_w$ is *total*.

certain worlds be "ruled out" as alternatives for φ. The "relevant" worlds in this case are the closest $\neg\varphi$-worlds, and these must be "ruled out" in the sense that the agent *does not believe* φ in these worlds. Hence we can think of the sensitivity condition

$$\mathrm{Min}_{\leqslant_w}\big(\overline{[\![\varphi]\!]}_h\big) \cap [\![B\varphi]\!]_h = \emptyset$$

as fitting into the pattern of

$$\mathsf{r}([\![\varphi]\!]^{\mathfrak{M}}, w) \cap \mathsf{u}([\![\varphi]\!]^{\mathfrak{M}}, w) = \emptyset.$$

Similarly, we can think of the *safety* condition as a requirement that certain worlds be "ruled out." The "relevant" worlds in this case are any $\neg\varphi$-worlds within the fixed set $\mathrm{Min}_{\leqslant_w}(W)$ of close worlds, and these must be "ruled out" in the sense that the agent does not believe φ in these worlds. Hence we can think of safety condition

$$\mathrm{Min}_{\leqslant_w}(W) \cap \overline{[\![\varphi]\!]}_s \cap [\![B\varphi]\!]_s = \emptyset.$$

as also fitting into the pattern of

$$\mathsf{r}([\![\varphi]\!]^{\mathfrak{M}}, w) \cap \mathsf{u}([\![\varphi]\!]^{\mathfrak{M}}, w) = \emptyset.$$

What this show is that there are two natural ways of transforming a given CB model $\mathcal{M} = \langle W, D, \leqslant, V\rangle$ into a SAB model $\mathfrak{M}_x = \langle W, D, \mathsf{u}, \mathsf{r}_x, V\rangle$, where $x \in \{h, s\}$:

$$\mathsf{u}(P, w) = \{v \in W \mid D(v) \subseteq P\}; \tag{3.25}$$

$$\mathsf{r}_h(P, w) = \mathrm{Min}_{\leqslant_w}(\overline{P}); \tag{3.26}$$

$$\mathsf{r}_s(P, w) = \mathrm{Min}_{\leqslant_w}(W) \cap \overline{P}. \tag{3.27}$$

We have already seen in §3.3.1 the properties of r implied by (3.26) and (3.27), which are the same as (3.21) and (3.22). Let us give the key property of u an official name.

Definition 3.16 (dox)**.** Given $\mathfrak{M} = \langle W, D, \mathsf{u}, \mathsf{r}, V\rangle$, u satisfies dox iff for all $w \in W$

and $P \subseteq W$,

$$\mathsf{u}(P, w) = \{v \in W \mid D(v) \subseteq P\}.$$

Hence for any formula φ,

$$\mathsf{u}(\llbracket\varphi\rrbracket^{\mathfrak{M}}, w) = \llbracket B\varphi\rrbracket^{\mathfrak{M}}.$$

In other words, dox says that we take the uneliminated alternatives for P to be the worlds in which the agent believes P. If one prefers to think of the uneliminated alternatives for P as not-P worlds, then we can instead take the uneliminated alternatives for P to be the not-P worlds in which the agent believes P, as follows.

Definition 3.17 (Fdox)**.** Given $\mathfrak{M} = \langle W, D, \mathsf{u}, \mathsf{r}, V\rangle$, u satisfies Fdox iff for all $w \in W$ and $P \subseteq W$,

$$\mathsf{u}(P, w) = \{v \in W \mid v \in \overline{P} \text{ and } D(v) \subseteq P\}.$$

Hence for any formula φ,

$$\mathsf{u}(\llbracket\varphi\rrbracket^{\mathfrak{M}}, w) = \overline{\llbracket\varphi\rrbracket}^{\mathfrak{M}} \cap \llbracket B\varphi\rrbracket^{\mathfrak{M}}.$$

Clearly both dox and Fdox give us the desired equivalences with H/S-semantics:

$$\begin{array}{lllllll}
\mathfrak{M}_h, w \vDash K\varphi & \text{iff} & D(w) \subseteq \llbracket\varphi\rrbracket & \text{and} & & & \\
 & & \mathsf{r}_d(\llbracket\varphi\rrbracket, w) & \cap & \mathsf{u}(\llbracket\varphi\rrbracket, w) & = & \emptyset \\
 & & \| & & \| & & \\
 & \text{(sensitivity)} & \mathrm{Min}_{\leqslant_w}\big(\overline{\llbracket\varphi\rrbracket}\big) & \cap & \overline{\llbracket\varphi\rrbracket} \cap \llbracket B\varphi\rrbracket & = & \emptyset; \\
 & & & & & & \\
\mathfrak{M}_s, w \vDash K\varphi & \text{iff} & D(w) \subseteq \llbracket\varphi\rrbracket & \text{and} & & & \\
 & & \mathsf{r}_l(\llbracket\varphi\rrbracket, w) & \cap & \mathsf{u}(\llbracket\varphi\rrbracket, w) & = & \emptyset \\
 & & \| & & \| & & \\
 & \text{(safety)} & \mathrm{Min}_{\leqslant_w}(W) \cap \overline{\llbracket\varphi\rrbracket} & \cap & \overline{\llbracket\varphi\rrbracket} \cap \llbracket B\varphi\rrbracket & = & \emptyset.
\end{array}$$

We are now ready to define the two special classes of SAB models.

Definition 3.18 (CB-like SAB Model Classes)**.**

1. Let H be the class of SAB models in which u satisfies dox/Fdox and r satisfies the same conditions as in model class D of Definition 3.13.3 (contrast, r-RofA, noVK, alpha, beta).

2. Let S be the class of SAB models in which u satisfies dox/Fdox and r satisfies the same conditions as in model class L of Definition 3.13.2 (r-RofA and $\mathsf{RS}_{\exists\forall}$).

Finally, we can state the result that we have been building up to: Theorem 3.2 shows that the H/S-semantics of §2 over CB models are equivalent, as semantics for the epistemic language, to our new semantics over SAB models in H/S. Therefore, the completeness theorems of Chapter 2 transfer to the SAB model classes.

Theorem 3.2 (CB and SAB Models)**.** Let $x \in \{h, s\}$ and $\mathsf{X} \in \{\mathsf{H}, \mathsf{S}\}$.

1. For every total, well-founded, CB model $\mathcal{N}$ with the universal field property, there is a SA model $\mathfrak{N}_x \in \mathsf{X}$ such that for all epistemic-doxastic formulas φ,

$$\mathcal{N}, w \vDash_x \varphi \text{ iff } \mathfrak{N}_x, w \vDash \varphi;$$

2. For every SAB model $\mathfrak{N} \in \mathsf{X}$, there is a total, well-founded CB model $\mathcal{N}$ with the universal field property such that for all epistemic-doxastic formulas φ,

$$\mathfrak{N}, w \vDash \varphi \text{ iff } \mathcal{N}, w \vDash_x \varphi.$$

The proof of this result is essentially the same as for Theorem 3.1 in Appendix 3.A.

Building in Belief

Before extending our analysis to N-semantics with the adherence condition, let us first consider further the interpretation of uneliminated possibilities in terms belief. The dox and Fdox conditions state a relationship between the u function and the doxastic accessibility relation D. A natural question is what these relationships imply about the properties of the u function by itself. The following provides the answer.

Observation 3.3 (Doxastic Conditions on u)**.**

1. Suppose a SA model $\langle W, \mathsf{u}, \mathsf{r}, V\rangle$ satisfies the following conditions:

   (a) $\mathsf{u}(P, w) = \mathsf{u}(P, v)$;

   (b) if $P \subseteq Q$, then $\mathsf{u}(P) \subseteq \mathsf{u}(Q)$;

   (c) for $\Sigma \subseteq \mathcal{P}(W)$, $\bigcap\limits_{P\in\Sigma} \mathsf{u}(P) \subseteq \mathsf{u}(\bigcap\limits_{P\in\Sigma} P)$;

   (d) $\bigcap\{P \subseteq W \mid w \in \mathsf{u}(P)\} \neq \emptyset$.

   Then if we define a binary relation D on W by

$$D(w) = \bigcap\{P \subseteq W \mid w \in \mathsf{u}(P)\}, \tag{3.28}$$

   $\mathfrak{M} = \langle W, D, \mathsf{u}, \mathsf{r}, V\rangle$ is a SAB model satisfying dox.

2. If a SAB model $\mathfrak{M} = \langle W, D, \mathsf{u}, \mathsf{r}, V\rangle$ satisfies dox, then u satisfies (a) - (d).

Before giving the simple proof of this observation, let us consider it conceptually. Recall that the notion of elimination used in §3.3.1 is *world-relative*: it is allowed that in world w, the agent has not eliminated possibility x as an alternative for P, while in world v, the agent has eliminated possibility x as an alternative for P. By contrast, condition (a) says that the notion of elimination implied by dox is not world-relative, but rather "global": x is uneliminated by the agent in w as an alternative for P just in case the agent believes P in x, so w drops out. However, the notion of elimination implied by dox is obviously *proposition-relative*. Condition (b) says that if x is uneliminated as an alternative for some proposition, then x is uneliminated as an alternative for any weaker proposition; condition (c) says that if x is uneliminated as an alternative for each of the propositions in Σ, then x is uneliminated as an alternative for the conjunction of these propositions; and condition (d) says that the set of proposition with respect to which x is uneliminated is consistent. It is easy to check that any u function satisfying dox satisfies conditions (a) - (d), which clearly reflect our idealized model of fully-closed belief for an ideally astute logician.

Moreover, if u satisfies these conditions, then we can define the set of worlds $D(w)$ compatible with what the agent believes in w to be the set of worlds in which all of

the propositions with respect to which w is uneliminated are true (or by (c), in which the strongest such proposition is true), and the dox relationship will hold.

Proof. For part 1, the seriality of D follows from (3.28) and (d), so $\mathfrak{M}$ is indeed a SAB model. Now we must show that for all $Q \subseteq W$,

$$\mathsf{dox} \quad w \in \mathsf{u}(Q) \text{ iff } D(w) \subseteq Q.$$

The left-to-right direction is immediate from (3.28). For the right-to-left direction, it follows from (d) that

$$w \in \bigcap_{P \in \{P \subseteq W \mid w \in \mathsf{u}(P)\}} \mathsf{u}(P), \tag{3.29}$$

which with (c) and (3.28) implies

$$w \in \mathsf{u}(\bigcap\{P \subseteq W \mid w \in \mathsf{u}(P)\}) = \mathsf{u}(D(w)). \tag{3.30}$$

It follows from $D(w) \subseteq Q$ and (b) together that $\mathsf{u}(D(w)) \subseteq \mathsf{u}(Q)$, which with (3.30) implies $w \in \mathsf{u}(Q)$. Part 2 is also straightforward. □

An analogous observation applies in the case of Fdox, which delivers the result that the possibilities uneliminated as alternatives for P are not-P possibilities.

Observation 3.4 (Doxastic Conditions on u cont.)**.**

1. Suppose a SA model $\langle W, \mathsf{u}, \mathsf{r}, V\rangle$ satisfies the following conditions:

   u-contrast $\quad \mathsf{u}(P, w) \subseteq \overline{P}$;

   (a′) $\quad \mathsf{u}(P, w) = \mathsf{u}(P, v)$;

   (b′) $\quad$ if $P \subseteq Q$, then $\mathsf{u}(P) \cap \overline{Q} \subseteq \mathsf{u}(Q)$;

   (c′) $\quad$ for $\Sigma \subseteq \mathcal{P}(W)$, $\bigcap\limits_{P \in \Sigma} \mathsf{u}(P) \subseteq \mathsf{u}(\bigcap\limits_{P \in \Sigma} P)$;

   (d′) $\quad$ if $\{P \subseteq W \mid w \in \mathsf{u}(P)\} \neq \emptyset$, then $\bigcap\{P \subseteq W \mid w \in \mathsf{u}(P)\} \neq \emptyset$.

Then if we define a binary relation D on W by

$$D(w) = \begin{cases} \bigcap\{P \subseteq W \mid w \in \mathsf{u}(P)\} & \text{if this is non-empty} \\ \{w\} & \text{otherwise} \end{cases}, \tag{3.31}$$

$\mathfrak{M} = \langle W, D, \mathsf{u}, \mathsf{r}, V\rangle$ is a SAB model satisfying Fdox.

2. If a SAB model $\mathfrak{M} = \langle W, D, \mathsf{u}, \mathsf{r}, V\rangle$ satisfies Fdox, then u satisfies u-contrast and (a′) - (d′).

Proof. For part 1, the seriality of D follows from (3.31), so $\mathfrak{M}$ is indeed a SAB model. Now we must show that for all $Q \subseteq W$,

$$\textsf{Fdox} \quad w \in \mathsf{u}(Q) \text{ iff } w \in \overline{Q} \text{ and } D(w) \subseteq Q.$$

The left-to-right direction is immediate from u-contrast, (d′), and (3.31). For the right-to-left direction, given $w \in \overline{Q}$ and $D(w) \subseteq Q$, we have $D(w) \neq \{w\}$, which with (3.31) implies

$$w \in \bigcap_{P \in \{P \subseteq W \mid w \in \mathsf{u}(P)\}} \mathsf{u}(P), \tag{3.32}$$

which with (c′), (3.31), and $D(w) \neq \{w\}$ implies

$$w \in \mathsf{u}(\bigcap\{P \subseteq W \mid w \in \mathsf{u}(P)) = \mathsf{u}(D(w)). \tag{3.33}$$

It follows from $D(w) \subseteq Q$ and (b′) together that $\mathsf{u}(D(w)) \cap \overline{Q} \subseteq \mathsf{u}(Q)$, which with (3.33) and $w \in \overline{Q}$ implies $w \in \mathsf{u}(Q)$. Part 2 is also straightforward. □

While our focus in this chapter has been on the "theory of the r function," Observations 3.3 and 3.4 concern the "theory of the u function" when we interpret elimination in terms of the beliefs of an ideally astute logician. In Chapter 5, I will give a different theory of the u function, according to which elimination is both world- and proposition-relative, in order to model the role of deduction in extending knowledge.

Adding Adherence

Our final step in unification is to incorporate Nozick's full tracking theory, as formalized by the N-semantics of §2.5. The question is how we are to understand satisfying the adherence condition in terms of ruling out possibilities. In fact, to capture adherence we must generalize our interpretation of the r and u functions.

Like the other conditions on knowledge that we have considered, the adherence condition requires that the agent fulfill some *epistemic success condition* with respect to some *selected set of possibilities*. For sensitivity with respect to φ, the selected worlds are the closest $\neg\varphi$-worlds, and the epistemic success conditions is *not believing* φ in these worlds. For safety with respect to φ, the selected worlds are any $\neg\varphi$-worlds within the fixed set $\mathrm{Min}_{\leqslant_w}(W)$ of close worlds, and the epistemic success condition is again *not believing* φ in these worlds. By contrast, for adherence with respect to φ, the selected worlds are any φ-worlds within the fixed set of $\mathrm{Min}_{\leqslant_w}(W)$ of close worlds, and the epistemic success condition is *believing* φ in these worlds. In each case, knowledge requires that the selected set of possibilities—$\mathsf{r}(P, w)$—does not overlap with the set of possibilities with respect to which the epistemic success condition is *unfulfilled*—$\mathsf{u}(P, w)$. To combine more than one of these conditions in a single theory, we simply distinguish different pairs of r and u functions for each condition, as follows.

Definition 3.19 (SAB$\times n$ Model)**.** A SAB$\times n$ model is a tuple $\mathfrak{M}$ of the form $\langle W, D, \{\mathsf{u}_i\}_{i\leq n}, \{\mathsf{r}_i\}_{i\leq n}, V\rangle$ where for all $i \leq n$, $\langle W, D, \mathsf{u}_i, \mathsf{r}_i, V\rangle$ is a SAB model.

The truth clause for $K\varphi$ is as before, but now with n distinct "no overlap" conditions.

Definition 3.20 (Truth in a SAB$\times n$ Model)**.** Given a SAB$\times n$ model $\mathfrak{M} = \langle W, D, \{\mathsf{u}_i\}_{i\leq n}, \{\mathsf{r}_i\}_{i\leq n}, V\rangle$ with $w \in W$ and a formula φ in the epistemic language, we define $\mathfrak{M}, w \vDash \varphi$ as follows (with other cases as before):

$$\mathfrak{M}, w \vDash K\varphi \quad \text{iff} \quad D(w) \subseteq [\![\varphi]\!]^{\mathfrak{M}} \text{ and } \forall i \leq n\colon \mathsf{r}_i([\![\varphi]\!]^{\mathfrak{M}}, w) \cap \mathsf{u}_i([\![\varphi]\!]^{\mathfrak{M}}, w) = \emptyset.$$

Let us think of r_1 and u_1 as the functions for the sensitivity condition, and r_2 and u_2 as the functions for the adherence condition. Now we can define a class of SAB models that captures the tracking theory as formalized by N-semantics.

Definition 3.21 (track)**.** A SAB×2 model $\langle W, D, \{\mathsf{u}_1, \mathsf{u}_2\}, \{\mathsf{r}_1, \mathsf{r}_2\}, V\rangle$ satisfies track if and only if:

1. r_1 satisfies contrast, r-RofA, noVK, alpha, and beta;
2. $\mathsf{r}_2(P, w) = \mathsf{r}_1(\emptyset, w) \cap P$;
3. $\mathsf{u}_1(P, w) = \{v \in W \mid D(v) \subseteq P\}$ (dox)
   (or $\mathsf{u}_1(P, w) = \{v \in W \mid v \in \overline{P} \text{ and } D(v) \subseteq P\}$ (Fdox));
4. $\mathsf{u}_2(P, w) = \{v \in W \mid D(v) \not\subseteq P\}$
   (or $\mathsf{u}_2(P, w) = \{v \in W \mid v \in P \text{ and } D(v) \not\subseteq P\}$).

We have already discussed the conditions on r_1 and u_1 in the previous sections. The condition on u_2 says that for adherence to P, the worlds with respect to which the epistemic success condition is unfulfilled by the agent are the worlds in which (P is true but) the agent does not believe P. Finally, consider the equation $\mathsf{r}_2(P, w) = \mathsf{r}_1(\emptyset, w) \cap P$. As we have seen, we can think of r_1 as encoding rankings of worlds, such that $\mathsf{r}_1(P, w) = \mathrm{Min}_{\leqslant_w}(\overline{P})$. As a special case, we have $\mathsf{r}_1(\emptyset, w) = \mathrm{Min}_{\leqslant_w}(\overline{\emptyset}) = \mathrm{Min}_{\leqslant_w}(W)$, so we can think of $\mathsf{r}_1(\emptyset, w)$ as the set of worlds closest to w. If we take the selected worlds for adherence to P to be the P-worlds among the closest worlds, $\mathrm{Min}_{\leqslant_w}(W) \cap P$ as in N-semantics, then this means $\mathsf{r}_2(P, w) = \mathsf{r}_1(\emptyset, w) \cap P$.

With this explanation, we can state an analogue of Theorem 3.2 for N-semantics.

Theorem 3.3 (CB and SAB×2 Models)**.**

1. For every total, well-founded CB model $\mathcal{N}$ with the universal field property, there is a SAB×2 model $\mathfrak{N}$ satisfying track such that for all epistemic-doxastic formulas φ,
$$\mathcal{N}, w \vDash_n \varphi \text{ iff } \mathfrak{N}, w \vDash \varphi;$$
2. For every SAB×2 model $\mathfrak{N}$ satisfying track, there is a total, well-founded CB model $\mathcal{N}$ with the universal field property such that for all epistemic-doxastic formulas φ,
$$\mathfrak{N}, w \vDash \varphi \text{ iff } \mathcal{N}, w \vDash_n \varphi.$$

The proof of this result is similar to that of Theorems 3.2 and 3.1.

Remark 3.2 (Free-Floating Adherence)**.** In the SAB×2 framework, it is clear that there is nothing preventing a tracking theorist from proposing that the set of selected worlds for adherence to P is not equal to the set of P-worlds *among the closest worlds according to the ranking for sensitivity*, but rather the set of selected worlds for adherence to P may float free of the ranking for sensitivity. For example, one might replace Definition 3.21.2 with the condition on r_2 that for all $w \in W$,

$$\text{there is } (\exists)\ \mathsf{R}_2(w) \subseteq W \text{ such that for all } (\forall)\ P \subseteq W,\ \mathsf{r}_2(P, w) = \mathsf{R}_2(w) \cap P.$$

Definition 3.21.2 implies this $\exists\forall$ condition (recall §3.2.1), but not vice versa. With the weaker condition, one could interpret $\mathsf{R}_2(w)$ as the set of "close" or "nearby"—rather than closest—worlds (cf. Heller 1989, Heller 1999a, and Pritchard 2005, 72). Both versions of the tracking theory fit neatly into the SAB×2 framework. It is easy to see that so does DeRose's [2004] double-safety theory combining safety and adherence.

Finally, since the SAB×2 framework does not require the existence of a fixed adherence sphere $\mathsf{R}_2(w)$ as above, it can also capture the $\forall\exists$ version of adherence suggested by Nozick [1981, 680n8], as mentioned in note 24 of Chapter 2.

With Theorems 3.1 - 3.3, unification is complete. The C/D/L- and H/N/S-semantics of Chapters 2 are indeed special cases of Fallibilism 1.0.

3.4 Conclusion

As promised at the beginning of this chapter, we have developed a unifying framework into which all of the RA and subjunctivist pictures in Chapters 2 fit as special cases. Doing so has illuminated the structural assumptions built into these pictures, as well as the relation of these assumptions to the closure properties of knowledge. This perspective will play a crucial role in Chapters 4 and 5, allowing us to clearly see the flaws of Fallibilism 1.0 and the path to a new framework of Fallibilism 2.0.

3.A Unification: Proofs

In this appendix, we prove Theorem 3.1 of §3.3.1.

Theorem 3.1 (RA and SA Models)**.** Let $x \in \{c, d, l\}$ and $\mathsf{X} \in \{\mathsf{C}, \mathsf{D}, \mathsf{L}\}$.

1. For every total, well-founded RA model $\mathcal{M}$ with the universal field property,[11] there is a SA model $\mathfrak{M}_x \in \mathsf{X}$ such that for all epistemic formulas φ,

$$\mathcal{M}, w \vDash_x \varphi \text{ iff } \mathfrak{M}_x, w \vDash \varphi.$$

2. For every SA model $\mathfrak{M} \in \mathsf{X}$, there is a total, well-founded RA model $\mathcal{M}$ with the universal field property such that for all epistemic formulas φ,

$$\mathfrak{M}, w \vDash \varphi \text{ iff } \mathcal{M}, w \vDash_x \varphi.$$

Proof. For part 1, given the RA model $\mathcal{M} = \langle W, \rightarrowtriangle, \preceq, V\rangle$, define the SA model $\mathfrak{M}_x = \langle W, \mathsf{u}, \mathsf{r}_x, V\rangle$ with

$$\begin{aligned} \mathsf{u}(P, w) &= \rightarrowtriangle(w) \cap \overline{P}; && (3.34)\\ \mathsf{r}_c(P, w) &= \overline{P}; \\ \mathsf{r}_d(P, w) &= \mathrm{Min}_{\preceq_w}(\overline{P}); && (3.35)\\ \mathsf{r}_l(P, w) &= \mathrm{Min}_{\preceq_w}(W) \cap \overline{P}. \end{aligned}$$

We will prove that $\mathfrak{M}_d \in \mathsf{D}$, leaving it to the reader to check that $\mathfrak{M}_c \in \mathsf{C}$ and $\mathfrak{M}_l \in \mathsf{L}$.

First we must show that u satisfies the following properties:

$\mathsf{RO}_{\exists\forall}$	there is $\mathsf{U}(w) \subseteq W$ such that for all $P \subseteq W$, $\mathsf{u}(P, w) = \mathsf{U}(w) \cap \overline{P}$.
u-RofA	if $w \notin P$, then $w \in \mathsf{u}(P, w)$.

That u satisfies $\mathsf{RO}_{\exists\forall}$ follows from (3.34) with $\mathsf{U}(w) = \rightarrowtriangle(w)$; that u satisfies u-RofA follows from the **reflexivity** of $\rightarrowtriangle$. Next we must show that r_d satisfies the following:

[11] Recall from Definition 2.3 that the universal field property requires that the field W_w of each $\preceq_w$ be W.

contrast	$\mathsf{r}(P,w) \subseteq \overline{P}$;
r-RofA	if $w \notin P$, then $w \in \mathsf{r}(P,w)$;
noVK	if $P \neq W$, then $\mathsf{r}(P,w) \neq \emptyset$;
alpha	$\mathsf{r}(P \cap Q, w) \subseteq \mathsf{r}(P,w) \cup \mathsf{r}(Q,w)$;
beta	if $P \subseteq Q$ and $\mathsf{r}(P,w) \cap \mathsf{r}(Q,w) \neq \emptyset$, then $\mathsf{r}(Q,w) \subseteq \mathsf{r}(P,w)$.

That r_d satisfies contrast is immediate from (3.35); that r_d satisfies r-RofA follows given the **weak centering** (Definition 2.2.3b) of $\preceq_w$; that r_d satisfies noVK follows given the **well-foundedness** and **universal field** of $\preceq_w$. To show that r_d satisfies alpha, we show that

$$\mathrm{Min}_{\preceq_w}(\overline{P \cap Q}) \subseteq \mathrm{Min}_{\preceq_w}(\overline{P}) \cup \mathrm{Min}_{\preceq_w}(\overline{Q}).$$

Suppose $v \in \mathrm{Min}_{\preceq_w}(\overline{P \cap Q}) = \mathrm{Min}_{\preceq_w}(\overline{P} \cup \overline{Q})$. Consider the case where $v \in \overline{P}$, and suppose for *reductio* that there is some $x \in \overline{P}$ such that $x \prec_w v$. Then since $x \in \overline{P} \cup \overline{Q}$, it follows that $v \notin \mathrm{Min}_{\preceq_w}(\overline{P} \cup \overline{Q})$, a contradiction.[12] Hence there is no such x, which means $v \in \mathrm{Min}_{\preceq_w}(\overline{P})$. The case where $v \in \overline{Q}$ is analogous.

Finally, to show that r_d satisfies beta, we show that

$$\text{if } P \subseteq Q \text{ and } \mathrm{Min}_{\preceq_w}(\overline{P}) \cap \mathrm{Min}_{\preceq_w}(\overline{Q}) \neq \emptyset, \text{ then } \mathrm{Min}_{\preceq_w}(\overline{Q}) \subseteq \mathrm{Min}_{\preceq_w}(\overline{P}).$$

Assume the antecedent and $v \in \mathrm{Min}_{\preceq_w}(\overline{Q})$, so $v \in \overline{P}$ given $P \subseteq Q$. Suppose for *reductio* that there is some $x \in \overline{P}$ such that $x \prec_w v$. By assumption, there is some $y \in \mathrm{Min}_{\preceq_w}(\overline{P}) \cap \mathrm{Min}_{\preceq_w}(\overline{Q})$. It follows by the **totality** of $\preceq_w$ that $y \preceq_w x$. By the **transitivity** of $\preceq_w$, $y \preceq_w x$ and $x \prec_w v$ implies $y \prec_w v$. Since $y \in \overline{Q}$, it follows that $v \notin \mathrm{Min}_{\preceq_w}(\overline{Q})$, a contradiction. Hence there is no such x, so $v \in \mathrm{Min}_{\preceq_w}(\overline{P})$.[13]

We conclude that $\mathfrak{M}_d \in \mathsf{D}$. The proof that $\mathcal{M}, w \Vdash_d \varphi$ iff $\mathfrak{M}_d, w \Vdash \varphi$ is by a straightforward induction on φ, using the truth definitions, (3.34), and (3.35).

For part 2, given a SA model $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle \in \mathsf{D}$, we construct the RA model

[12] Recall that $\mathrm{Min}_{\preceq_w}(S) = \{v \in S \cap W_w \mid \text{there is no } u \in S \text{ such that } u \prec_w v\}$.

[13] Note that this proof shows that r_d satisfies Bordes's [1976, §2] stronger condition β^+ below.

$\mathcal{M} = \langle W, \twoheadrightarrow, \preceq, V\rangle$ by defining[14]

$$x \twoheadrightarrow y \quad \text{iff} \quad y \in \mathsf{U}(x); \tag{3.37}$$

$$x \preceq_w y \quad \text{iff} \quad \exists Q \subseteq W\colon x \in \mathsf{r}(Q, w) \text{ and } y \in \overline{Q}. \tag{3.38}$$

We will show that $\preceq_w$ is a well-founded, total preorder with a universal field, weakly centered around w, and

$$\mathsf{r}(P, w) = \mathrm{Min}_{\preceq_w}(\overline{P}). \tag{3.39}$$

Then the fact that $\mathfrak{M}, w \vDash \varphi$ iff $\mathcal{M}, w \vDash_d \varphi$ follows by a straightforward induction.

To show: $\preceq_w$ is **total**.[15] Suppose for *reductio* that there are $x, y \in W$ such that $x \not\preceq_w y$ and $y \not\preceq_w x$. Hence by (3.38) there is no Q such that either $x \in \mathsf{r}(Q, w)$ and $y \in \overline{Q}$, or $y \in \mathsf{r}(Q, w)$ and $x \in \overline{Q}$. In particular, $\overline{\{x, y\}}$ is not such a Q, so given $x, y \in \{x, y\}$, we have $x \notin \mathsf{r}(\overline{\{x, y\}}, w)$ and $y \notin \mathsf{r}(\overline{\{x, y\}}, w)$. Then since $\mathsf{r}(\overline{\{x, y\}}, w) \subseteq \{x, y\}$ given $\mathsf{contrast}$, we have $\mathsf{r}(\overline{\{x, y\}}, w) = \emptyset$. It follows by noVK that $\{x, y\} = W$. Now consider $\mathsf{r}(\emptyset, w)$. Given $\emptyset \neq W$, we have $\mathsf{r}(\emptyset, w) \neq \emptyset$ by noVK again. Then since $W = \{x, y\}$, either $x \in \mathsf{r}(\emptyset, w)$ or $y \in \mathsf{r}(\emptyset, w)$. But $x, y \in \overline{\emptyset}$, so $\emptyset$ is a Q as specified above, contradicting the assumption for *reductio*.

[14] Given $\mathsf{contrast}$, noVK, alpha and beta, (3.38) is equivalent to the following definition (see Sen 1971, Definitions 2 and 5, Theorem T.3):

$$x \preceq_w y \text{ iff } x \in \mathsf{r}(\overline{\{x, y\}}, w).$$

Note how this definition takes advantage of the fact that the first input to the r function is a proposition, rather than a formula. If the first input to r were a formula, as in the SSA models of §4.B, then there would be no guarantee that $\overline{\{x, y\}}$ is definable by a formula. A definition that works even if the first input to r must be definable is the following:

$$x \preceq_w y \text{ iff either (i) } \forall P : y \notin \mathsf{r}(P, w) \text{ or (ii) } \exists Q : x \in \mathsf{r}(Q, w) \text{ and } y \in \overline{Q}. \tag{3.36}$$

[15] Here is a proof of totality using (3.36) from note 14. Suppose for *reductio* that there are $x, y \in W$ such that $x \not\preceq_w y$ and $y \not\preceq_w x$. Hence by (3.36) we have that (a) there are P and P' such that $x \in \mathsf{r}(P, w)$ and $y \in \mathsf{r}(P', w)$, but (b) there is no Q such that either $x \in \mathsf{r}(Q, w)$ and $y \in \overline{Q}$, or $y \in \mathsf{r}(Q, w)$ and $x \in \overline{Q}$. Given $x \in \mathsf{r}(P, w)$ and $y \in \mathsf{r}(P', w)$, it follows by $\mathsf{contrast}$ that $x \in \overline{P}$ and $y \in \overline{P'}$, so $x, y \in \overline{P \cap P'}$, which with (b) implies that $x \notin \mathsf{r}(P \cap P', w)$ and $y \notin \mathsf{r}(P \cap P', w)$. But by Observation 3.2.1, together alpha and beta imply that either $\mathsf{r}(P, w) \subseteq \mathsf{r}(P \cap P', w)$ or $\mathsf{r}(P', w) \subseteq \mathsf{r}(P \cap P', w)$, which with the fact that $x \in \mathsf{r}(P, w)$ and $y \in \mathsf{r}(P', w)$ implies that either $x \in \mathsf{r}(P \cap P', w)$ or $y \in \mathsf{r}(P \cap P', w)$, contradicting what was just derived.

To show: $\preceq_w$ has a **universal field**. Immediate since $\preceq_w$ is total on W.

To show: $\preceq_w$ is **transitive**. Suppose $x \preceq_w y$ and $y \preceq_w z$, so by (3.38) there are $P, Q \subseteq W$ such that $x \in \mathsf{r}(P, w)$, $y \in \overline{P} \cap \mathsf{r}(Q, w)$, and $z \in \overline{Q}$. Where $S = P \cap Q$, we have $z \in \overline{S}$, so if we can show that $x \in \mathsf{r}(S, w)$, then we will have $x \preceq_w z$ by (3.38).

Following Bordes [1976], we use the fact that alpha and beta together imply[16]

$$\beta^+ \quad \text{if } X \subseteq Y \text{ and } \mathsf{r}(X, w) \cap \overline{Y} \neq \emptyset, \text{ then } \mathsf{r}(Y, w) \subseteq \mathsf{r}(X, w).$$

Suppose for *reductio* that $\mathsf{r}(S, w) \cap \overline{P} = \emptyset$. Given contrast, $\mathsf{r}(S, w) \subseteq \overline{P \cap Q} = \overline{P} \cup \overline{Q}$ and $x \in \mathsf{r}(P, w) \subseteq \overline{P} \subseteq \overline{S}$, so $S \neq W$ and therefore $\mathsf{r}(S, w) \neq \emptyset$ by noVK. Hence from $\mathsf{r}(S, w) \cap \overline{P} = \emptyset$ we have $\mathsf{r}(S, w) \cap \overline{Q} \neq \emptyset$, which with β^+ implies $\mathsf{r}(Q, w) \subseteq \mathsf{r}(S, w)$. But then since $y \in \mathsf{r}(Q, w) \cap \overline{P}$, we have $y \in \mathsf{r}(S, w) \cap \overline{P}$, contradicting the assumption for *reductio*. Then given $\mathsf{r}(S, w) \cap \overline{P} \neq \emptyset$, it follows by β^+ that $\mathsf{r}(P, w) \subseteq \mathsf{r}(P \cap Q, w)$, so $x \in \mathsf{r}(S, w)$. Hence $x \preceq_w z$ as desired.

To show: (3.39) holds. For the left-to-right inclusion, if $x \in \mathsf{r}(P, w)$, then for any $y \in \overline{P}$, we have $x \preceq_w y$ by (3.38). It follows by the totality and transitivity of $\preceq_w$ that there is no $y \in \overline{P}$ such that $y \prec_w x$. Hence $x \in \mathrm{Min}_{\preceq_w}(\overline{P})$. From left-to-right, suppose $x \in \mathrm{Min}_{\preceq_w}(\overline{P})$. Hence $P \neq W$, so $\mathsf{r}(P, w) \neq \emptyset$ by noVK. Consider some $y \in \mathsf{r}(P, w)$. Given contrast, $y \in \overline{P}$, so $x \in \mathrm{Min}_{\preceq_w}(\overline{P})$ implies $x \preceq_w y$ by the totality and transitivity of $\preceq_w$. Hence by (3.38) there is some $Q \subseteq W$ such that $x \in \mathsf{r}(Q, w)$ and $y \in \overline{Q}$. Given $x \in \mathrm{Min}_{\preceq_w}(\overline{P})$, $x \in \overline{P}$, but suppose for *reductio* that $x \notin \mathsf{r}(P, w)$. Now recall from Remark 3.1 that alpha and beta are together equivalent to Arrow:

$$\text{if } X \subseteq Y \text{ and } \mathsf{r}(X, w) \cap \overline{Y} \neq \emptyset, \text{ then } \mathsf{r}(Y, w) = \mathsf{r}(X, w) \cap \overline{Y}.$$

Given $P, Q \subseteq P \cup Q$, $y \in \mathsf{r}(P, w) \cap \overline{P \cup Q}$, and $x \in \mathsf{r}(Q, w) \cap \overline{P \cup Q}$, Arrow implies

$$\mathsf{r}(P \cup Q) = \mathsf{r}(P, w) \cap \overline{P \cup Q} = \mathsf{r}(Q, w) \cap \overline{P \cup Q},$$

which contradicts the combination of $x \notin \mathsf{r}(P, w) \cap \overline{P \cup Q}$ and $x \in \mathsf{r}(Q, w) \cap \overline{P \cup Q}$.

[16] If $X \subseteq Y$, then $\mathsf{r}(X, w) \cap \overline{Y} \subseteq \mathsf{r}(Y, w)$ by the equivalent form of alpha given in Observation 3.1. Putting this together with $\mathsf{r}(X, w) \cap \overline{Y} \neq \emptyset$, we have $\mathsf{r}(X, w) \cap \mathsf{r}(Y, w) \neq \emptyset$, which is the antecedent of beta, which has the same consequent as β^+.

Hence $x \in \mathsf{r}(P, w)$.

To show: $\preceq_w$ is **well-founded**. Immediate from (3.39) and noVK.

To show: $\preceq_w$ is **weakly-centered**. Immediate from (3.39) and r-RofA. □

Remark 3.3 (Dropping Totality)**.** Theorems 3.1 - 3.3 are stated for *total* preorders. To see how to define classes of SA models corresponding to RA and CB models without the assumption of totality, recall from Remark 3.1 that the alpha and beta conditions are together equivalent to the Arrow condition:

$$\text{if } X \subseteq Y \text{ and } \mathsf{r}(X, w) \cap \overline{Y} \neq \emptyset, \text{ then } \mathsf{r}(Y, w) = \mathsf{r}(X, w) \cap \overline{Y}.$$

The Arrow condition is in turn equivalent to what is known in the economics literature as the Weak Axiom of Revealed Preference (rewritten using our r instead of c):

$$\mathsf{WARP} \quad \text{if } v \in \overline{P} \text{ and } \exists u \in \mathsf{r}(P, w)\colon v \preceq_w u, \text{ then } v \in \mathsf{r}(P, w),$$

where $\preceq_w$ is defined from r as in (3.38). If r satisfies WARP, contrast, and noVK, then $\preceq_w$ is a total preorder and $\mathsf{r}(P, w) = \mathrm{Min}_{\preceq_w}(\overline{P})$; and for any total preorder $\leq_w$ on W, if we define r by $\mathsf{r}(P, w) = \mathrm{Min}_{\leq_w}(\overline{P})$, then r satisfies WARP and contrast.[17] To obtain the analogous result without totality, Eliaz and Ok [2006] identify a weakening of WARP that they call the Weak Axiom of Revealed Non-Inferiority:

$$\mathsf{WARNI} \quad \text{if } v \in \overline{P} \text{ and } \forall u \in \mathsf{r}(P, w)\colon v \preceq_w u, \text{ then } v \in \mathsf{r}(P, w),$$

where $\preceq_w$ is defined from r as in (3.38). It follows from results of Eliaz and Ok that if r satisfies WARNI, contrast, and noVK, then we can define a preorder $\leq_w$ such that $\mathsf{r}(P, w) = \mathrm{Min}_{\leq_w}(\overline{P})$;[18] and for any preorder $\leq_w$ on W (where W is assumed to be countable), if we define r by $\mathsf{r}(P, w) = \mathrm{Min}_{\leq_w}(\overline{P})$, then r satisfies WARNI and contrast.

Remark 3.4 (Dropping Universality)**.** Theorems 3.1 - 3.3 are stated for preorders with the *universal field property* requiring that the field of each $\preceq_w$ is W. To restate

[17] For noVK, we must add the assumption that $\leq_w$ is well-founded.

[18] Eliaz and Ok show several ways of defining a preorder, other than the definition of $\preceq_w$ in (3.38), which differ in the further properties they guarantee (see their Remark 2 and Proof of Theorem 2).

the theorems without this assumption, we must associate with each $w \in W$ in our SA models a set $W_w \subseteq W$ and reformulate our conditions on r accordingly:

contrast*	$\mathsf{r}(P,w) \subseteq \overline{P} \cap W_w$;
r-RofA	if $w \notin P$, then $w \in \mathsf{r}(P,w)$;
noVK*	if $W_w \not\subseteq P$, then $\mathsf{r}(P,w) \neq \emptyset$;
alpha	$\mathsf{r}(P \cap Q, w) \subseteq \mathsf{r}(P,w) \cup \mathsf{r}(Q,w)$;
beta*	if $P \cap W_w \subseteq Q$ and $\mathsf{r}(P,w) \cap \mathsf{r}(Q,w) \neq \emptyset$, then $\mathsf{r}(Q,w) \subseteq \mathsf{r}(P,w)$.

Now everything works as before, with W_w serving as the field of $\preceq_w$.

3.B Relation to Neighborhood Models

To prove the completeness results of Propositions 3.1 - 3.6, our strategy is to relate our SA models to *neighborhood* models (see Chellas 1980, Ch. 9).

Definition 3.22 (Neighborhood Model)**.** A *neighborhood model* is a tuple $\mathbb{M} = \langle W, N, V\rangle$ where W and V are as in Definition 3.1 and $N\colon W \to \mathcal{P}(\mathcal{P}(W))$.

Definition 3.23 (Truth in a Neighborhood Model)**.** Given a neighborhood model $\mathbb{M} = \langle W, N, V\rangle$, we define $\mathbb{M}, w \vDash \varphi$ as follows (with propositional cases as usual):

$$\mathbb{M}, w \vDash K\varphi \text{ iff } [\![\varphi]\!]^{\mathbb{M}} \in N(w),$$

where $[\![\varphi]\!]^{\mathbb{M}} = \{v \in W \mid \mathbb{M}, v \vDash \varphi\}$.

We will show that every neighborhood model satisfying certain properties can be transformed into a modally equivalent SA model satisfying corresponding properties, and vice versa. We can then transfer completeness results for the classes of neighborhood models to completeness results for the corresponding classes of SA models.

Lemma 3.1 (SA Models and Neighborhood Models)**.** Suppose $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$ is a SA model satisfying some of the following three conditions:[19]

[19] As before, universal quantification over $w \in W$ and $P, Q \subseteq W$ is implicit.

1. $\mathsf{r}(P, w) \subseteq \overline{P}$ (contrast);

2. if $w \notin P$, then $w \in \mathsf{r}(P, w)$ and $w \in \mathsf{u}(P, w)$ (r-RofA and u-RofA);

3. if $P \subseteq Q$, then $\mathsf{r}(Q, w) \subseteq \mathsf{r}(P, w)$ (cover); and there is $\mathsf{U}(w) \subseteq W$ such that for all $P \subseteq W$, $\mathsf{u}(P, w) = \mathsf{U}(w) \cap \overline{P}$ ($\mathsf{RO}_{\exists\forall}$).

Then there is a neighborhood model $\mathbb{M} = \langle W, N, V\rangle$ satisfying the corresponding conditions among the following (i.e., 1 corresponds to 4, 2 to 5, and 3 to 6):

4. $W \in N(w)$ (contains the unit);

5. If $P \in N(w)$, then $w \in P$;

6. If $P \subseteq Q$, then $P \in N(w)$ implies $Q \in N(w)$ (supplemented);

and for all φ, $\mathfrak{M}, w \vDash \varphi$ iff $\mathbb{M}, w \vDash \varphi$.

In the other direction, if there is a neighborhood model $\mathbb{M} = \langle W, N, V\rangle$ satisfying some of 4 - 6, then there is a SA model $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$ satisfying the corresponding conditions among 1 - 3, such that for all φ, $\mathfrak{M}, w \vDash \varphi$ iff $\mathbb{M}, w \vDash \varphi$.

Proof. Given the SA model $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$, we construct a neighborhood model $\mathbb{M} = \langle W, N, V\rangle$ as follows. For all $w \in W$,

$$N(w) = \{P \subseteq W \mid \mathsf{r}(P, w) \cap \mathsf{u}(P, w) = \emptyset\}. \tag{3.40}$$

Consider properties 4 - 6 above for $\mathbb{M}$:

(iv) For condition 4, it follows from condition 1 (contrast) that $\mathsf{r}(W, w) = \emptyset$, which implies $W \in N(w)$ by (3.40).

(v) For condition 5, if $P \in N(w)$, then $\mathsf{r}(P, w) \cap \mathsf{u}(P, w) = \emptyset$ by (3.40). But if $w \notin P$, then $w \in \mathsf{r}(P, w) \cap \mathsf{u}(P, w) \neq \emptyset$ by condition 2 (r-RofA and u-RofA). Hence $w \in P$.

(vi) For condition 6, suppose $P \subseteq Q$ and $P \in N(w)$, so $\mathsf{r}(P,w) \cap \mathsf{u}(P,w) = \emptyset$ by (3.40). By the first part of condition 3 (cover), $P \subseteq Q$ implies $\mathsf{r}(Q,w) \subseteq \mathsf{r}(P,w)$, and by the second part ($\mathsf{RO}_{\exists\forall}$), $P \subseteq Q$ implies

$$\mathsf{u}(Q,w) = \mathsf{U}(w) \cap \overline{Q} \subseteq \mathsf{U}(w) \cap \overline{P} = \mathsf{u}(P,w).$$

It follows that $\mathsf{r}(Q,w) \cap \mathsf{u}(Q,w) = \emptyset$. Hence $Q \in N(w)$ by (3.40).

Having checked conditions 4 - 6 for $\mathbb{M}$, the proof that $\mathfrak{M}, w \vDash \varphi$ iff $\mathbb{M}, w \vDash \varphi$ is a straightforward induction on φ.

In the other direction, given a neighborhood model $\mathbb{M} = \langle W, N, V\rangle$, we construct a SA model $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$ as follows. For all $w \in W$ and $P \subseteq W$,

$$\mathsf{u}(P,w) = \overline{P}, \tag{3.41}$$

so u satisfies $\mathsf{RO}_{\exists\forall}$ and r-RofA by construction, and

$$\mathsf{r}(P,w) = \begin{cases} \emptyset & \text{if } P \in N(w) \\ \overline{P} & \text{if } P \notin N(w). \end{cases} \tag{3.42}$$

Consider conditions 1 - 3 above for $\mathfrak{M}$:

(i) Condition 1 (contrast) is immediate from (3.42).

(ii) For condition 2 (r-RofA and u-RofA), if $w \notin P$, then $P \notin N(w)$ by condition 5, so $w \in \mathsf{r}(P,w) = \overline{P}$ by (3.42), and $w \in \mathsf{u}(P,w) = \overline{P}$ by (3.41).

(iii) For condition 3, the second part ($\mathsf{RO}_{\exists\forall}$) is immediate from (3.41). For the first part (cover), assume $P \subseteq Q$. Case 1: $P \notin N(w)$, so $\mathsf{r}(P,w) = \overline{P}$ by (3.42). Since $P \subseteq Q$, $\overline{Q} \subseteq \overline{P}$, and by (3.42), $\mathsf{r}(Q,w) \subseteq \overline{Q}$. Hence $\mathsf{r}(Q,w) \subseteq \mathsf{r}(P,w)$, as desired. Case 2: $P \in N(w)$, so by condition 6, $Q \in N(w)$. It follows by (3.42) that $\mathsf{r}(Q,w) = \emptyset$, in which case $\mathsf{r}(Q,w) \subseteq \mathsf{r}(P,w)$ again.

Having checked conditions 1 - 3 for $\mathfrak{M}$, the proof that $\mathfrak{M}, w \vDash \varphi$ iff $\mathbb{M}, w \vDash \varphi$ is another straightforward induction on φ. □

Lemma 3.2 (SA and Neighborhood Models II)**.** Suppose $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$ is a SA model satisfying the following conditions:

1. $\mathsf{r}(P, w) \subseteq \overline{P}$ (contrast);
2. if $w \notin P$, then $w \in \mathsf{r}(P, w)$ and $w \in \mathsf{u}(P, w)$ (r-RofA and u-RofA);
3. if $P \subseteq Q$, then $\mathsf{r}(P, w) \cap \overline{Q} \subseteq \mathsf{r}(Q, w)$ (alpha) (recall Observation 3.1); and there is $\mathsf{U}(w) \subseteq W$ such that for all $P \subseteq W$, $\mathsf{u}(P, w) = \mathsf{U}(w) \cap \overline{P}$ ($\mathsf{RO}_{\exists\forall}$).

Then there is a neighborhood model $\mathbb{M} = \langle W, N, V\rangle$ satisfying the following:

4. $W \in N(w)$ (contains the unit);
5. If $P \in N(w)$, then $w \in P$;
6. If $\Sigma \subseteq N(w)$, then $\bigcap \Sigma \in N(w)$ (closed under intersection);

and for all φ, $\mathfrak{M}, w \vDash \varphi$ iff $\mathbb{M}, w \vDash \varphi$.

In the other direction, if there is a neighborhood model $\mathbb{M} = \langle W, N, V\rangle$ satisfying conditions 4 - 6, then there is a SA model $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$ satisfying conditions 1 - 3, such that for all φ, $\mathfrak{M}, w \vDash \varphi$ iff $\mathbb{M}, w \vDash \varphi$.

Proof. Given the SA model $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$, we construct a neighborhood model $\mathbb{M} = \langle W, N, V\rangle$ as before. For all $w \in W$,

$$N(w) = \{P \subseteq W \mid \mathsf{r}(P, w) \cap \mathsf{u}(P, w) = \emptyset\}. \tag{3.43}$$

We have already checked in the proof of Lemma 3.1 that (3.43) implies that conditions 4 and 5 hold for $\mathbb{M}$. Let us now check that the new condition 6 holds for $\mathbb{M}$. If $\Sigma \subseteq N(w)$, then we have the following three facts. First, for all $P \in \Sigma$, we have $\mathsf{r}(P, w) \cap \mathsf{u}(P, w) = \emptyset$ by (3.43). Second, for all $P \in \Sigma$, since $\bigcap \Sigma \subseteq P$ we have $\mathsf{r}(\bigcap \Sigma, w) \cap \overline{P} \subseteq \mathsf{r}(P, w)$ by alpha. Third, we have $\mathsf{u}(\bigcap \Sigma, w) = \mathsf{U}(w) \cap \overline{\bigcap \Sigma}$ by $\mathsf{RO}_{\exists\forall}$. Now suppose for *reductio* that there is some $v \in \mathsf{r}(\bigcap \Sigma, w) \cap \mathsf{u}(\bigcap \Sigma, w)$, which with the third fact above implies that $v \in \mathsf{U}(w) \cap \overline{Q}$ for some $Q \in \Sigma$, so $v \in \mathsf{u}(Q, w)$ by $\mathsf{RO}_{\exists\forall}$. Given $v \in \overline{Q}$, it follows by the second fact above that $v \in \mathsf{r}(Q, w)$ as well.

Thus, we have $\mathsf{r}(Q,w) \cap \mathsf{u}(Q,w) \neq \emptyset$, contradicting the first fact above. We conclude that $\mathsf{r}(\bigcap \Sigma, w) \cap \mathsf{u}(\bigcap \Sigma, w) = \emptyset$, so $\bigcap \Sigma \in N(w)$ by (3.43), as desired.

In the other direction, given a neighborhood model $\mathbb{M} = \langle W, N, V\rangle$, we construct a SA model $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$ as follows. For every $P \notin N(w)$, consider the set

$$S_P^w = \bigcap \{Q \subseteq W \mid P \subseteq Q \in N(w)\}.$$

By condition 5, $w \in S_P^w$. Given $P \subseteq S_P^w$, if $\overline{P} \cap S_P^w = \emptyset$, then $P = S_P^w$, in which case $P \in N(w)$ by condition 6, contradicting our assumption. Therefore,

$$\overline{P} \cap S_P^w \neq \emptyset.$$

For each $P \notin N(w)$, choose an element $u_P^w \in \overline{P} \cap S_P^w$ such that if $w \notin P$, then $u_P^w = w$. Next, define[20]

$$\mathsf{U}(w) = \{u_P^w \mid P \notin N(w)\}, \tag{3.44}$$

and for all $Q \subseteq W$, let

$$\mathsf{u}(Q,w) = \mathsf{U}(w) \cap \overline{Q}. \tag{3.45}$$

Observe that u satisfies $\mathsf{RO}_{\exists\forall}$ and u-RofA.

Finally, we define r as follows:

$$\mathsf{r}(Q,w) = \{u_P^w \mid P \subseteq Q \text{ and } P \notin N(w)\} \cap \overline{Q}. \tag{3.46}$$

Clearly r satisfies contrast. For r-RofA, if $w \notin P$, then $P \notin N(w)$ by condition 5, in which case $u_P^w = w$ by the construction above, so $w \in \mathsf{r}(P,w)$ by (3.46). Finally, it is immediate from (3.46) that r satisfies alpha, written in the alternative form given in Observation 3.1 as: if $P \subseteq Q$, then $\mathsf{r}(P,w) \cap \overline{Q} \subseteq \mathsf{r}(Q,w)$.

It only remains to show that $\mathfrak{M}, w \vDash \varphi$ iff $\mathbb{M}, w \vDash \varphi$.

If $\mathbb{M}, w \nvDash K\varphi$, then $[\![\varphi]\!]^{\mathbb{M}} \notin N(w)$. It follows by (3.44) - (3.46) that

[20] I am assuming the Axiom of Choice, but this is not necessary. We could define $U_P^w = \overline{P} \cap S_P^w$ and $\mathsf{U}(w) = \bigcup_{P \notin N(w)} U_P^w$, adjusting the rest of the proof accordingly. In any case, for reasons indicated in note 21 below, we only need this direction of the lemma for finite $\mathbb{M}$.

$$u^w_{[\![\varphi]\!]^{\mathbb{M}}} \in \mathsf{r}([\![\varphi]\!]^{\mathbb{M}}, w) \cap \mathsf{u}([\![\varphi]\!]^{\mathbb{M}}, w)$$

and hence

$$u^w_{[\![\varphi]\!]^{\mathfrak{M}}} \in \mathsf{r}([\![\varphi]\!]^{\mathfrak{M}}, w) \cap \mathsf{u}([\![\varphi]\!]^{\mathfrak{M}}, w)$$

by the inductive hypothesis, so $\mathfrak{M}, w \not\models K\varphi$.

If $\mathbb{M}, w \vDash K\varphi$, then $[\![\varphi]\!]^{\mathbb{M}} \in N(w)$ and hence $[\![\varphi]\!]^{\mathfrak{M}} \in N(w)$ by the inductive hypothesis. By (3.46),

$$\mathsf{r}([\![\varphi]\!]^{\mathfrak{M}}, w) = \{u^w_P \mid P \subseteq [\![\varphi]\!]^{\mathfrak{M}} \text{ and } P \notin N(w)\} \cap \overline{[\![\varphi]\!]}^{\mathfrak{M}},$$

so if there is some $u \in \mathsf{r}([\![\varphi]\!]^{\mathfrak{M}}, w)$, then $u = u^w_P$ for some $P \subseteq [\![\varphi]\!]^{\mathfrak{M}}$. Recall that

$$u^w_P \in S^w_P = \bigcap \{Q \subseteq W \mid P \subseteq Q \in N(w)\}.$$

Then since $P \subseteq [\![\varphi]\!]^{\mathfrak{M}} \in N(w)$, it follows that $u^w_P \in [\![\varphi]\!]^{\mathfrak{M}}$, contradicting the fact that $u^w_P \in \mathsf{r}([\![\varphi]\!]^{\mathfrak{M}}, w) \subseteq \overline{[\![\varphi]\!]}^{\mathfrak{M}}$. Hence $\mathsf{r}([\![\varphi]\!]^{\mathfrak{M}}, w) = \emptyset$, which implies $\mathfrak{M}, w \vDash K\varphi$. □

Lemma 3.3 (SA and Neighborhood Models III)**.** Suppose $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$ is a SA model satisfying the following conditions:

1. $\mathsf{r}(P, w) \subseteq \overline{P}$ (contrast);
2. if $w \notin P$, then $w \in \mathsf{r}(P, w)$ and $w \in \mathsf{u}(P, w)$ (r-RofA and u-RofA);
3. there is $\mathsf{R}(w) \subseteq W$ such that for all $P \subseteq W$, $\mathsf{r}(P, w) = \mathsf{R}(w) \cap \overline{P}$ ($\mathsf{RS}_{\exists\forall}$); and there is $\mathsf{U}(w) \subseteq W$ such that for all $P \subseteq W$, $\mathsf{u}(P, w) = \mathsf{U}(w) \cap \overline{P}$ ($\mathsf{RO}_{\exists\forall}$).

Then there is a neighborhood model $\mathbb{M} = \langle W, N, V\rangle$ satisfying the following:

4. $W \in N(w)$ (contains the unit);
5. If $P \in N(w)$, then $w \in P$;
6. $P \in N(w)$ iff $\bigcap N(w) \subseteq P$ (augmented);

and for all φ, $\mathfrak{M}, w \vDash \varphi$ iff $\mathbb{M}, w \vDash \varphi$.

In the other direction, if there is a neighborhood model $\mathbb{M} = \langle W, N, V\rangle$ satisfying conditions 4 - 6, then there is a SA model $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$ satisfying conditions 1 - 3, such that for all φ, $\mathfrak{M}, w \vDash \varphi$ iff $\mathbb{M}, w \vDash \varphi$.

Proof. Given the SA model $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$, we construct a neighborhood model $\mathbb{M} = \langle W, N, V\rangle$ as before. For all $w \in W$,

$$N(w) = \{P \subseteq W \mid \mathsf{r}(P, w) \cap \mathsf{u}(P, w) = \emptyset\}. \tag{3.47}$$

We already checked in the proof of Lemma 3.1 that (3.47) implies conditions 4 and 5 for $\mathbb{M}$. Let us now check the new condition 6 for $\mathbb{M}$. As noted with (3.1) in §3.2.1, assuming $\mathsf{RS}_{\exists\forall}$ and $\mathsf{RO}_{\exists\forall}$, $\mathsf{r}(P, w) \cap \mathsf{u}(P, w) = \emptyset$ is equivalent to $\mathsf{R}(w) \cap \mathsf{U}(w) \subseteq P$. Thus, by (3.47), $P \in N(w)$ is equivalent to $\mathsf{R}(w) \cap \mathsf{U}(w) \subseteq P$; moreover $\bigcap N(w) = \bigcap\{Q \subseteq W \mid \mathsf{R}(w) \cap \mathsf{U}(w) \subseteq Q\} = \mathsf{R}(w) \cap \mathsf{U}(w)$, so $P \in N(w)$ iff $\bigcap N(w) \subseteq P$.

In the other direction, given a neighborhood model $\mathbb{M} = \langle W, N, V\rangle$, we construct a SA model $\mathfrak{M} = \langle W, \mathsf{u}, \mathsf{r}, V\rangle$ as follows:

$$\begin{aligned} \mathsf{r}(P, w) &= \overline{P}; \\ \mathsf{U}(w) &= \bigcap N(w); \\ \mathsf{u}(P, w) &= \mathsf{U}(w) \cap \overline{P}. \end{aligned}$$

It is easy to see that $\mathfrak{M}$ satisfies conditions 1 - 3 and that $\mathfrak{M}, w \vDash \varphi$ iff $\mathbb{M}, w \vDash \varphi$. □

As explained in §3.2, condition 1/4 in Lemmas 3.1 - 3.3 corresponds to the necessitation rule N (or the axiom $K\top$); condition 2/5 in Lemmas 3.1 - 3.3 corresponds to the T axiom, $K\varphi \to \varphi$; in Lemma 3.1, condition 3/6 corresponds to the M axiom, $K(\varphi \wedge \psi) \to K\varphi \wedge K\psi$; in Lemma 3.2, condition 3/6 corresponds to the C axiom, $K\varphi \wedge K\psi \to K(\varphi \wedge \psi)$; and in Lemma 3.3, condition 3/6 corresponds to the K axiom, $(K\varphi \wedge K(\varphi \to \psi)) \to K\psi$. For proofs that the systems combining these axioms in Propositions 3.1 - 3.6 (**E**, **EN**, **ENT**, **K**, **KT**, **EMNT**, **ECNT**) are sound and complete for the appropriate classes of neighborhood models, see Chellas 1980, Ch.

9.[21] Since Lemma 3.1 shows that any formula falsified by a neighborhood model in a certain class is also falsified by a SA model in the corresponding class, and vice versa, the cited completeness results for the classes of neighborhood models transfer to the completeness results in Propositions 3.1 - 3.6 for SA models.

[21]The condition corresponding to the C axiom in neighborhood models is closure under *finite* intersections, whereas I have stated Lemma 3.2 using closure under arbitrary intersections. Here we use the fact that ECNT is complete with respect to the class of finite neighborhood models satisfying conditions 4 - 6 of Lemma 3.2 [Chellas, 1980, §9.5], for which the two intersection conditions coincide.

4

The Flaws of Fallibilism 1.0

Having built up the framework of Fallibilism 1.0, we will now evaluate it critically. As suggested at the beginning of Chapter 3, I will argue that any way of navigating down the tree of Fig. 4.1 leads to one of three serious problems:

- The Problem of Vacuous Knowledge;
- The Problem of Containment;
- The Problem of Knowledge Inflation.

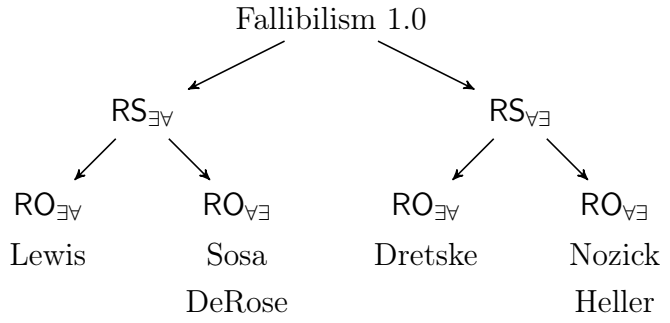

Figure 4.1: theories classified by RS and RO parameter settings

After explaining the first two problems in §4.1 and §4.2, in §4.2.1 I will prove an impossibility result to the effect that they are unavoidable by any version of Fallibilism

1.0. In §4.3, I will consider and reject an attempt to escape the impossibility result by "knowledge inflation." My conclusion will be that Fallibilism 1.0 is inherently flawed. However, this negative point leads in a positive direction, as we apply what we have learned about the flaws of Fallibilism 1.0 to develop Fallibilism 2.0 in Chapter 5.

4.1 The Problem of Vacuous Knowledge

The Problem of Vacuous Knowledge arises for any fallibilist theory with a $\mathsf{RS}_{\exists\forall}$ parameter setting—any theory that goes left starting from Fallibilism 1.0 in Fig. 4.2.

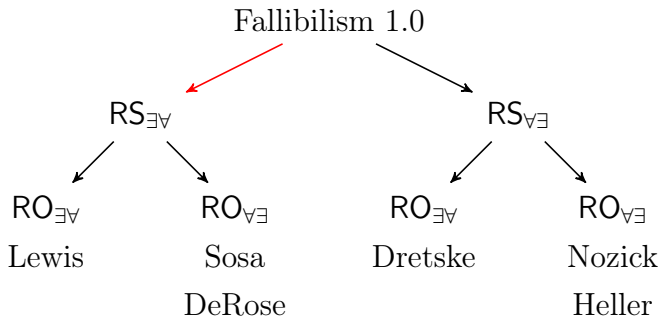

Figure 4.2: parameter settings and the Problem of Vacuous Knowledge

Recall that on a $\mathsf{RS}_{\exists\forall}$ theory, for any context $\mathcal{C}$ and world w, there is a fixed set $\mathsf{R}_{\mathcal{C}}(w) \subseteq W$ of worlds such that for any proposition P, knowing P in w relative to $\mathcal{C}$ requires ruling out a world v iff v is a not-P world in $\mathsf{R}_{\mathcal{C}}(w)$: $\mathsf{r}_{\mathcal{C}}(P, w) = \mathsf{R}_{\mathcal{C}}(w) \cap \overline{P}$. (I omit $\mathcal{C}$ when no confusion should arise.) As discussed in Chapter 2 and §3.2.1, RA theorists like Stine [1976] and Lewis [1996] who wish to maintain full closure assume $\mathsf{RS}_{\exists\forall}$. So do safety and double-safety theorists like Sosa [1999] and DeRose [2004], who nonetheless fail to preserve full closure due to the failure of $\mathsf{RO}_{\exists\forall}$ (see §4.2).

One must pay a serious price for assuming $\mathsf{RS}_{\exists\forall}$: either accept *infallibilism* or allow *vacuous knowledge*—knowledge of contingent propositions gained without the elimination of a single possibility. The problem can be illustrated very simply as in Fig. 4.3. Assuming $\mathsf{RS}_{\exists\forall}$, $\mathsf{fallibilism}$ says that the fixed set of "relevant" worlds $\mathsf{R}(w)$ is a strict subset of the whole space W. Consider any contingent proposition

$Q \neq W$ that is true throughout $\mathsf{R}(w)$, i.e., $\mathsf{R}(w) \subseteq Q$. Then given $\mathsf{RS}_{\exists\forall}$, we have $\mathsf{r}(Q,w) = \mathsf{R}(w) \cap \overline{Q} = \emptyset$, which violates the no vacuous knowledge condition:

$$\mathsf{noVK} \quad \text{if } Q \neq W, \text{ then } \mathsf{r}(Q,w) \neq \emptyset.$$

Hence even if the agent has not eliminated *any* possibilities, so $\mathsf{u}(P,w) = \overline{P}$ for all propositions $P \subseteq W$, nonetheless $\mathsf{r}(Q,w) \cap \mathsf{u}(Q,w) = \emptyset$, so the theory implies that the agent knows Q. Typical examples of such propositions Q, true throughout $\mathsf{R}(w)$ in ordinary contexts, are the denials of skeptical hypotheses. Hence a fallibilist $\mathsf{RS}_{\exists\forall}$ theory implies that agents can know the denials of skeptical hypotheses without having eliminating any possibilities of any kind, skeptical or mundane.

Fact 4.1 (VK Dilemma)**.** r satisfies $\mathsf{RS}_{\exists\forall}$ and noVK iff r satisfies infallibilism.

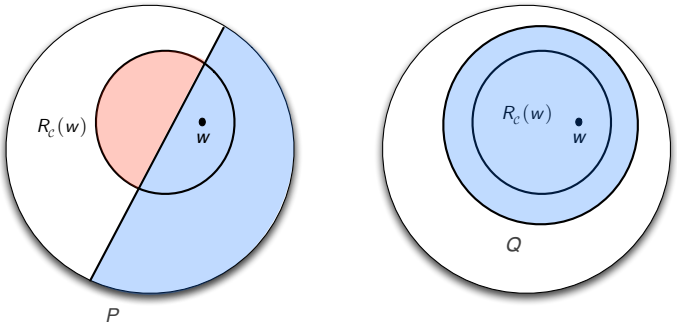

Figure 4.3: the Problem of Vacuous Knowledge illustrated

Together $\mathsf{RS}_{\exists\forall}$ and fallibilism violate the basic idea that knowledge of a contingent empirical proposition about the world is a cognitive achievement that does not come "for free" (or even cheap), in the sense of not requiring the elimination of any possibilities, but rather requires such epistemic work. This is what I call the Problem of Vacuous Knowledge, following the terminology of Heller [1999a, 207], who also realizes that the $\mathsf{RS}_{\exists\forall}$ assumption is to blame. However, Heller and I view the problem differently. For Heller, the problem seems to be that when a contingent Q is true throughout $\mathsf{R}(w)$, $\mathsf{RS}_{\exists\forall}$ theories do not place a requirement on the agent to eliminate any not-Q-possibilities in order to know Q. In my view, the problem is that $\mathsf{RS}_{\exists\forall}$

theories do not place on the agent *any requirement to eliminate any possibility* in order to know Q. This distinction will become important later in Chapter 5.

Of course, I make no pretense of having a proof or knockdown argument for noVK.[1] In fact, I suspect that the issue of noVK is a near-bedrock issue that marks a deep division among epistemologists. In the rest of this section, I will explain why I am unconvinced by attempts to make violations of noVK look acceptable.[2]

4.1.1 Reply 1: No Problem

The first reply to the Problem of Vacuous Knowledge is to claim that there is no problem with the idea of agents knowing contingent propositions without having eliminated any possibilities. About Dretske's zebra case, Stine [1976, 258] writes:

> [O]ne does know what one takes for granted in normal circumstances. I do know that it is not a mule painted to look like a zebra. I do not need evidence for such a proposition. The evidence picture of knowledge has been carried too far.... [I]f the negation of a proposition is not a relevant alternative, then I know it—obviously, without needing to provide evidence—and so obviously that it is odd, misleading even, to give utterance to my knowledge.

Of course, there is no argument here, but simply an assertion that "vacuous knowledge" is obviously knowledge. Against Stine's assertion, Cohen [1988, 99] writes:

> Here, I think Stine's strategy for preserving closure becomes strongly counter-intuitive. Even if it is true that some propositions can be known without evidence, surely this is not true of the proposition that S is not deceived by a cleverly disguised mule.

While this is just another assertion from Cohen, Stine's bald endorsement of vacuous knowledge is sufficiently controversial that we should consider other replies.

[1]Note added in ILLC version: in fact, in order to raise a problem for $\mathsf{RS}_{\exists\forall}$ theories, it is not necessary to argue that noVK holds for *all* propositions. See Holliday 2013b, §2.5.

[2]Note added in ILLC version: for updated and condensed versions of the arguments in the following sections, see Holliday 2013b, §2.4.

4.1.2 Reply 2: Unclaimable Knowledge

The second reply to the Problem of Vacuous Knowledge acknowledges that there is apparently a problem, but argues that it can be mitigated by appeal to *contextualism*. How can we try to smooth over the inconsistency between (a) $\mathsf{RS}_{\exists\forall}$, (b) the fallibilist position that for a mundane proposition P and a skeptical counter-hypothesis S, $\mathsf{r}_c(P, w)$ does not contain any skeptical S-possibilities, and (c) the noVK position that $\mathsf{r}_c(\overline{S}, w)$ should be non-empty? One way is to claim that when our conversation turns to the skeptical hypothesis S, we shift to a context $\mathcal{C}'$ in which $\mathsf{r}_{c'}(\overline{S}, w)$ is non-empty and $\mathsf{r}_{c'}(P, w)$ *does* contain S-possibilities, contrary to our initial fallibilist position but consistent with $\mathsf{RS}_{\exists\forall}$. This is exactly Lewis's [1996, 561-562] line:

> Do I claim you can know P just by presupposing it?! Do I claim you can know that a possibility W does not obtain just by ignoring it? Is that not what my analysis implies, provided that the presupposing and the ignoring are proper? Well, yes. And yet I do not claim it. Or rather, I do not claim it for any specified P or W. I have to grant, in general, that knowledge just by presupposing and ignoring *is* knowledge; but it is an *especially* elusive sort of knowledge, and consequently it is an unclaimable sort of knowledge. You do not even have to practise epistemology to make it vanish. Simply *mentioning* any particular case of this knowledge, aloud or even in silent thought, is a way to attend to the hitherto ignored possibility, and thereby render it no longer ignored, and thereby create a context in which it is no longer true to ascribe the knowledge in question to yourself or others. So, just as we should think, presuppositions alone are not a basis on which to *claim* knowledge.

According to Lewis, vacuous knowledge is unclaimable knowledge—or rather, claimable in the abstract, but not for any specific proposition. Does this resolve the problem?

I think not, for three reasons that I will sketch now and develop below:

1. The Mechanism Problem. Lewis's idea that simply mentioning or attending to a possibility changes the context in such a way that the possibility becomes

relevant has fallen out of favor with contemporary contextualists. According to DeRose [2004], we can resist such context changes, in which case what is stopping us from truly claiming specific instances of vacuous knowledge?

2. The Motivation Problem. What gets us into the Problem of Vacuous Knowledge is insistence on full closure. To maintain full closure, we need to know anti-skeptical propositions vacuously. Lewis says that vacuous knowledge is unclaimable knowledge. But then closure becomes *unclaimable closure*, so we have to ask ourselves about how we got into this mess: was it really worth it?

3. The Missing-the-Point Problem. As Cohen [2000] correctly frames the vacuous knowledge problem, it is that Lewis's theory (and Cohen's own) implies that agents have a special kind of contingent *a priori* knowledge. What is problematic about this is not just the idea that agents could *truly claim* to have such contingent *a priori* knowledge, but the idea that they could *have it* at all.

The Mechanism Problem

Lewis's [1996] claim, related to his *Rule of Attention* (559), is that when an agent vacuously knows a proposition P relative to an attributor's context $\mathcal{C}$, if the attributor says or thinks that the agent knows P, this will invariably change the attributor's context from $\mathcal{C}$ to a new $\mathcal{C}'$ in which not-P possibilities uneliminated by the agent are relevant. However, few contemporary contextualists think that sayings or thinkings invariably introduce relevant counter-possibilities in this way.[3] For example, DeRose [2004, Ch. 4] suggests that participants in a conversation may resist context changes by sticking to their own "personally indicated epistemic standards." In this case, resistance may be unnecessary. Imagine two lovers of vacuous knowledge, who delight in the ease with which they know various contingent empirical propositions without doing the epistemic work of eliminating any possibilities of error. Not wanting to leave this happy state, they are careful to stick to mutually indicated epistemic standards

[3]Cohen [1998, 303n24] suggests that the Rule of Attention may need to be defeasible; Ichikawa [2011, §4] disavows it; and Blome-Tillman [2009, 246-247] argues that it is too strong.

relative to which they have vacuous knowledge. According to post-Lewisian contextualism, there does not seem to be anything stopping them from truly claiming to know what they know vacuously. If this is correct, then the "unclaimable knowledge" reply collapses. But just to be safe, let us add two more problems with it.

The Motivation Problem

When Stine [1976] first articulated what I call the $\mathsf{RS}_{\exists\forall}$ condition, the motivation was clear: preserve closure for a fixed context. In those cases where Dretske claims that closure fails, Stine claims that fixed-context closure holds because our knowledge of the conclusion is vacuous (contrast this with the "knowledge inflation" of §4.3). If Stine saves fixed-context closure by appeal to vacuous knowledge, how significant is this result? Lewis's "unclaimable knowledge" reply threatens to make it rather insignificant. For if Lewis is correct and context change invariably prevents us from truly claiming to have the vacuous knowledge that is rightly ours according to fixed-context closure, why should we bend over backwards to preserve fixed-context closure?

As Dretske [2005] said of Lewis-style contextualism, "it is a way of preserving closure for the heavyweight implications while abandoning its usefulness in acquiring knowledge of them" (19). To put it this way is misleading, however, since a closure principle is not something that agents use in acquiring knowledge (unless we are talking about knowledge of what agents know). It is not as if in the course of a mathematical proof, the mathematician "uses closure" to extend her knowledge from premises to conclusion. Instead, a closure principle is something that we use in *reasoning about the knowledge of agents*—of others and ourselves. Hence Dretske's claim should be reformulated as: Lewis-style contextualism is a way of "preserving" closure while abandoning its usefulness in reasoning about knowledge.

In response to Dretske, Hawthorne [2005, 39] clarifies the contextualist view:

> Suppose that A isn't considering heavyweight possibilities but that B is. In particular suppose that B believes P and also goes on to believe some heavyweight consequence of P by deduction. Suppose A says (1) "B knows P" and, moreover, (2) "B knows anything he has deductively inferred from

> P and thereby come to believe." If B [sic] is in a context where "A [sic] knows P," in her mouth, expresses a truth, then she will be in a context where (2), in her mouth, expresses a truth as well.

(Note that A and B should be switched in the last sentence.) The trick, of course, is that A does not consider any specific heavyweight consequence Q of P. For if A were to do this, it could change her context so that it would *not* be true for her to say "B knows Q." This was Dretske's complaint, which Hawthorne does not address.

Insofar as Hawthorne's no-specificity trick allows contextualists to express their commitment to closure, it also allows us to state their commitment to vacuous knowledge. Indeed, Lewis stated his commitment to vacuous knowledge in exactly such an unspecific way, which leads to my last point about the "unclaimable knowledge" reply.

The Missing-the-Point Problem

The third problem with the "unclaimable knowledge" reply is that as a reply to the Problem of Vacuous Knowledge, it misses the point. What is problematic about vacuous knowledge is not just the idea that agents could *truly claim* to have it—which they probably can according to post-Lewisian contextualism—but rather that they could *have it* at all. As Cohen [2000, 105] concedes about his contextualist view:

> Unfortunately, a problem remains. If in everyday contexts, it is *a priori* rational enough for us to know the falsity of skeptical alternatives, then we have *a priori* knowledge of the falsity of skeptical alternatives. But surely these alternatives are contingent. So it looks as if the contextualist is committed to the view that we have contingent *a priori* knowledge. And of course, these cases do not fit the structure of the reference-fixing cases called to our attention by Kripke.
>
> Of course, I am not entirely happy with this result. . . .

I am not at all happy with non-Kripkean contingent *a priori* knowledge. Yet as I suggested before, I suspect that this is a near-bedrock issue that deeply divides epistemologists. For example, DeRose [2000, 138] is willing to say that we do have contingent *a priori* knowledge of the denials of skeptical hypotheses:

> I suspect that the best ways of filling out and then evaluating my alternative, contextualist Moorean account of how we know that we're not BIVs will have it come out also as an account according to which our knowledge that we're not BIVs is a priori.

What about skeptical hypotheses not about how our perception hooks up to the world, but rather about the constitution of the world around us (see the discussion of "world-side" skeptical hypotheses in §6.2.3)? Do we have *a priori* knowledge of their denials as well? I find such a view, given by fallibilist $\mathsf{RS}_{\exists\forall}$ theories, highly implausible. Without having a filled-out version of the contextualist Moorean account to evaluate, for now I will simply state that I am firmly in the camp of those who think that knowledge of contingent empirical propositions requires investigation of the world.[4]

4.1.3 Reply 3: Something Less Than Ruling Out

The third reply to the Problem of Vacuous Knowledge, considered (and rejected) by Vogel [1999], is to admit that knowledge of contingent truths always requires epistemic work, but to argue that this "epistemic work" may involve something less that *ruling out* any possibilities in a strong sense.[5] As Vogel [1999, 159 - 159n12] explains:

> The RAT is committed to the thesis that one can know that an irrelevant alternative is false even though one can't rule it out.... The RA theorist might still require that you have *some* minimal evidence against irrelevant alternatives in order to know that they are false. However, holding onto this scruple will make it more difficult, if not impossible, for the RA theorist to resist skepticism.

The presupposition of the last sentence seems to be that agents typically lack even minimal evidence against radical skeptical alternatives. Cohen [1988, 111] agrees:

[4]Note added in ILLC version (see Holliday 2013b, §2.4): even if one thinks there are some special counterexamples to Evans's [1979, 161] famous claim that "it would be intolerable for there to be a statement which is both knowable *a priori* and deeply contingent," such examples are beside the point. The point is that $\mathsf{RS}_{\exists\forall}$ theories imply that *every proposition* Q with $\mathsf{R}_c(w) \subseteq Q$ is knowable with no requirement of eliminating possibilities, and there is no guarantee that every such Q fits the mold of one of the recherché examples of (deeply) contingent but *a priori* knowable propositions.

[5]I am grateful to Zoltan Gendler Szabo for raising this idea in conversation.

> A moderate skeptical hypothesis is immune to rejection on the basis of a particular kind of evidence.... Radical skeptical hypotheses are immune to rejection on the basis of *any* evidence. There would appear to be no evidence that could count against the hypothesis that we are deceived by a Cartesian demon.... Radical skeptical hypotheses are designed to neutralize any evidence that could be adduced against them.

If this is correct, it will not help to hold onto $\mathsf{RS}_{\exists\forall}$ by saying that knowing P requires *ruling out* all not-P possibilities within $\mathsf{R}(w)$ (if any) and *having some evidence* against all not-P possibilities outside of $\mathsf{R}(w)$, since the second part will lead to skepticism. Followers of Williamson's [2000, Ch. 9] $E = K$ (evidence = knowledge) thesis may claim that we do have evidence against skeptical hypotheses insofar as we *know* they do not obtain. But then they owe us an explanation of what epistemic work we have done to earn such knowledge, or else we are back to the vacuous knowledge.[6]

It is worth emphasizing that the Problem of Vacuous Knowledge is as much a problem for *safety* theories [Sosa, 1999, Pritchard, 2005] as it is for $\mathsf{RS}_{\exists\forall}$ versions of the RA theory. One might claim that we have earned our anti-skeptical knowledge because our anti-skeptical beliefs are *safe*—there are no close worlds where the skeptical hypotheses are true but we believe they are false—but this is because they are *vacuously* safe—there are no close worlds where the skeptical hypotheses are true at all. The problem with safety in this case is precisely that it is a $\mathsf{RS}_{\exists\forall}$ condition.

4.1.4 Reply 4: Double-Safety

While safety suffers from the Problem of Vacuous Knowledge, what about the other $\exists\forall$ condition (recall Remark 3.2): *adherence*. The fourth reply to the problem is that for a given skeptical hypothesis S, even if one's anti-skeptical belief in not-S is

[6]Note added in ILLC version: even if Vogel and Cohen are incorrect and we can acquire at least minimal evidence against skeptical hypotheses, the "something less than ruling out" reply still does not solve the Problem of Vacuous Knowledge facing $\mathsf{RS}_{\exists\forall}$ theories. For while having "minimal evidence" may not require eliminating any $\neg P$-possibilities, where P is the contingent empirical proposition to be known, presumably it does requires eliminating *some* possibilities, perhaps as alternatives to related propositions (see Holliday 2013b). But if it does, then we must reject $\mathsf{RS}_{\exists\forall}$, since it allows agents to know contingent truths with *no* requirement of eliminating possibilities.

vacuously safe, in virtue of the fact that not-S is true throughout the set $\mathsf{R}(w)$ of nearby worlds, it is not vacuously *adherent*, since it is some kind of achievement that in all of the nearby worlds where not-S is true, the adherent agent believes not-S. As Heller [1999a, 207] puts it, "One might insist that [the agent] is in the following positive epistemic condition with respect to [not-S]: if p were true, [the agent] would believe p."[7] DeRose [2000, 135] suggests something similar:

> As a Moorean, I face the pointed question: *How* do we know that we're not BIVs? ... [M]y account is that we know, according to even quite high standards (though not according to the absolute standards) that we're not BIVs because our belief as to whether we're BIVs matches the fact of the matter in the actual world and in the sufficiently nearby worlds.

Does a "double-safety" theory (recall §2.10.1) combining safety and adherence solve the Problem of Vacuous Knowledge?[8]

Against the idea that adding adherence solves the Problem of Vacuous Knowledge, Heller objects that an agent "is in that condition [if p were true, she would believe p] with respect to any (or perhaps almost any) true belief" (207), perhaps because he thinks that a counterfactual with a true antecedent is equivalent to a material conditional. But if we take adherence to range over some set $\mathsf{R}(w)$ of nearby worlds, then it is not the case that an agent satisfies adherence with respect to any true belief. However, as I will explain, something similarly defeating is the case for double-safety.

As Kripke [2011, 183] shows, if an agent's belief that p satisfies the *sensitivity* condition, then normally the agent's belief that $p \wedge Bp$ will satisfy both the sensitivity and adherence conditions (see note 54 in Chapter 2). Kripke concludes that adherence "is almost without force, a broken reed. What can be the point of a condition whose rigor can almost always be overcome by conjoining 'and I believe (via M) that p'...?"' (184). A similar point applies to the adherence part of double-safety. Suppose an agent's belief that p is vacuously safe, in virtue of the fact that there are no $\neg p$-worlds among the nearby worlds. It follows by an argument similar to Kripke's that

[7] I have changed the variables in Heller's sentence for consistency with mine.

[8] I am grateful to Keith DeRose for raising this idea in conversation.

the agent's belief that $p \wedge Bp$ will normally be double-safe, even if the agent's belief that p is not. So according to the double-safety theory, it is normally sufficient to know $p \wedge Bp$ that one has a vacuously safe belief that p. I conclude that double-safety does not solve the Problem of Vacuous Knowledge, but only relocates it.

To put the point more formally, in the framework of S-semantics for safety in §2.5 and R-semantics for double-safety in §2.10.1, the following is easy to prove if we constrain the doxastic accessibility relation D such that $B\varphi \rightarrow BB\varphi$ is valid.

Fact 4.2 (Double-Safe Belief about Belief)**.** For any CB model $\mathcal{M} = \langle W, D, \leqslant, V\rangle$ in which D is transitive, $w \in W$, and propositional formula φ, $\mathcal{M}, w \Vdash_s K\varphi$ implies $\mathcal{M}, w \Vdash_r K(\varphi \wedge B\varphi)$.

Proof. Since D is transitive, $B\varphi \rightarrow BB\varphi$ is valid. To show $\mathcal{M}, w \Vdash_r K(\varphi \wedge B\varphi)$, we first observe that for any $v \in \mathrm{Min}_{\leqslant_w}(W)$, we have these equivalences: $\mathcal{M}, v \vDash B\varphi$ iff $\mathcal{M}, v \vDash B\varphi \wedge BB\varphi$ (because $B\varphi \rightarrow BB\varphi$ is valid) iff $\mathcal{M}, v \vDash B(\varphi \wedge B\varphi)$ (because $B\alpha \wedge B\beta \leftrightarrow B(\alpha \wedge \beta)$ is valid). Now given $\mathcal{M}, w \Vdash_s K\varphi$, we have $\mathcal{M}, w \Vdash_s B\varphi$, which for propositional φ implies $\mathcal{M}, w \Vdash_r B\varphi$, which with the previous equivalences implies $\mathcal{M}, w \Vdash_r B(\varphi \wedge B\varphi)$, so the belief condition for $\mathcal{M}, w \Vdash_r K(\varphi \wedge B\varphi)$ is satisfied. Moreover, given $\mathcal{M}, w \Vdash_s K\varphi$, we have that for all $v \in \mathrm{Min}_{\leqslant_w}(W)$, $\mathcal{M}, v \vDash B\varphi$ iff $\mathcal{M}, v \vDash \varphi \wedge B\varphi$. Thus, by the previous equivalences, $\mathcal{M}, v \vDash B(\varphi \wedge B\varphi)$ iff $\mathcal{M}, v \vDash \varphi \wedge B\varphi$, so the double-safety condition for $\mathcal{M}, w \Vdash_r K(\varphi \wedge B\varphi)$ is satisfied. □

4.1.5 $\mathsf{RS}_{\exists\forall}$ Reconsidered

Having found no satisfactory solution to the Problem of Vacuous Knowledge, fallibilists would be wise to reconsider the $\mathsf{RS}_{\exists\forall}$ assumption. What, after all, are the *arguments for* $\mathsf{RS}_{\exists\forall}$? It is not obvious that for every context $\mathcal{C}$ there is such a special set $\mathsf{R}_{\mathcal{C}}(w)$ as required by $\mathsf{RS}_{\exists\forall}$. Why should there be a single set for which the equation $\mathsf{r}_{\mathcal{C}}(P, w) = \mathsf{R}_{\mathcal{C}}(w) \cap \overline{P}$ holds for all propositions P, no matter how different the subject matters of these propositions? That there is such a single set is a substantial theoretical assumption. Infallibilism implies $\mathsf{RS}_{\exists\forall}$, where $\mathsf{R}_{\mathcal{C}}(w)$ is the set of *all* possibilities. Yet those who reject infallibilism but want to retain full closure must add $\mathsf{RS}_{\exists\forall}$ as a further assumption. What is the argument for this assumption?

As far as I know, the only person to attempt to argue for $\mathsf{RS}_{\exists\forall}$ directly was Stine [1976]. There are two parts to Stine's argument. First, Stine claims that to reject full closure by flouting $\mathsf{RS}_{\exists\forall}$, as Dretske did, "would be to commit some logical sin akin to equivocation" (256). This seems to be a confusion, repeated by those who endorse the equivocation charge [Cohen, 1988, 98]. There is no equivocation in the explanations in §3.2.1 and Chapter 2 of how closure can fail without $\mathsf{RS}_{\exists\forall}$. And to reject the additional $\mathsf{RS}_{\exists\forall}$ assumption on r is not equivocation either.

The second part of Stine's argument is more interesting. Concerning the validity of arguments involving 'knows', Stine claims that "If the relevant alternatives, which have after all to do with the truth or falsity of the premises and conclusion, cannot be held fixed," as in $\mathsf{RS}_{\exists\forall}$, then "it is hard to see on what basis one can decide whether the argument form is valid or not" (256). Must we choose between $\mathsf{RS}_{\exists\forall}$ and an anything-goes epistemic logic? The results of Chapter 3 demonstrate that the answer is 'no'. As we saw, fallibilists who reject $\mathsf{RS}_{\exists\forall}$ can accept a variety of other constraints on r, and given such a set of constraints, we can completely characterize the forms of valid argument involving 'knows'. Stine's second argument for $\mathsf{RS}_{\exists\forall}$ fails as well.

4.2 The Problem of Containment

As shown in §4.1, if we go left in the $\mathsf{RS}_{\exists\forall}$ direction down the tree in Fig. 4.2, we encounter the Problem of Vacuous Knowledge. A natural question is whether we might do better if we instead go right in the $\mathsf{RS}_{\forall\exists}$ direction down the tree in Fig. 4.4. Heller [1999b, 128n5] explains how a desire to avoid vacuous knowledge is what drives him away from the $\mathsf{RS}_{\exists\forall}$ condition of *safety* to the $\mathsf{RS}_{\forall\exists}$ condition of *sensitivity*:

> The property in question is the property of not believing p in any of the not-p worlds within the selected set. The simple version of the anti-luck theory would hold that the selected worlds are all and only the similar enough ones. However, when not-p is very bizarre there may not be any not-p worlds among the similar enough worlds. In such cases the simple version would be forced to attribute vacuous knowledge. To avoid this

> consequence, I prefer a more complicated version of the anti-luck theory according to which the selected worlds are all those that are close enough plus all those that are as close as the closest not-p worlds. This extra clause will only make a difference in cases in which there are no not-p worlds among the close enough worlds.... The simple version of the anti-luck condition preserves closure, while my more complicated version rejects closure.

As shown in Chapter 3, the properties of the r function built in to the world-ordering picture for sensitivity are the following:

contrast	$\mathsf{r}(P,w) \subseteq \overline{P}$;
r-RofA	if $w \notin P$, then $w \in \mathsf{r}(P,w)$;
noVK	if $P \neq W$, then $\mathsf{r}(P,w) \neq \emptyset$;
alpha	$\mathsf{r}(P \cap Q, w) \subseteq \mathsf{r}(P,w) \cup \mathsf{r}(Q,w)$;
beta	if $P \subseteq Q$ and $\mathsf{r}(P,w) \cap \mathsf{r}(Q,w) \neq \emptyset$, then $\mathsf{r}(Q,w) \subseteq \mathsf{r}(P,w)$.

The guarantee of noVK is an attractive feature of this package of properties. However, other pieces of the package create another serious problem: the Problem of Containment. We have already investigated this problem in detail in Chapters 2. Recall from the Closure Theorem of Chapter 2 that even such weak closure principles as the following fail for Heller and Nozick's theories of knowledge:[9]

- $K(\varphi \wedge \psi) \rightarrow K\varphi \wedge K\psi$;
- $K(\varphi \wedge \psi) \rightarrow K(\varphi \vee \psi)$;
- $K\varphi \wedge K\psi \rightarrow K(\varphi \vee \psi)$.

For reasons explained in Chapter 2, I regard such closure failures and their consequences (e.g., for higher-order knowledge) as unacceptable. Using the framework of Chapter 3, we can pinpoint the source of these closure failures. For Heller and Nozick,

[9]Subject to the qualification of Remark 2.4 for Heller.

the problem is that their theories do not satisfy the condition

$$\mathsf{cover} \quad \text{if } P \subseteq Q, \text{ then } \mathsf{r}(Q, w) \subseteq \mathsf{r}(P, w).$$

As we saw in §3.2.3, cover corresponds in a precise sense to $K(\varphi \wedge \psi) \rightarrow K\varphi \wedge K\psi$ in the framework of Chapter 3. Recall our intuitive reading of cover: if P excludes as much of logical space as Q does, then coming to know P should require at least as much epistemic work, in terms of eliminating possibilities, as coming to know Q does. For example, since $\varphi \wedge \psi$ excludes as much of logical space as $\varphi \vee \psi$ does, coming to know $\varphi \wedge \psi$ should require at least as much epistemic work, in terms of ruling out possibilities, as coming to know $\varphi \vee \psi$ does. Imagine if someone were to say, "I agree that you've done enough research to know that $\varphi \wedge \psi$, but when it comes to knowing $\varphi \vee \psi$, that's going to take some more work in the lab." This seems absurd, but if Heller and Nozick's theories were correct, it would make perfect sense: knowing $\varphi \wedge \psi$ only requires ruling out the closest $(\neg\varphi \vee \neg\psi)$-worlds, which may all be easy-to-eliminate $\neg\varphi$-worlds, whereas knowing $\varphi \vee \psi$ requires ruling out the closest $(\neg\varphi \wedge \neg\psi)$-worlds, and $\neg\psi$-worlds may be very difficult to eliminate.

What this shows is that the difference between cover and

$$\mathsf{beta} \quad \text{if } P \subseteq Q \text{ and } \mathsf{r}(P, w) \cap \mathsf{r}(Q, w) \neq \emptyset, \text{ then } \mathsf{r}(Q, w) \subseteq \mathsf{r}(P, w),$$

which the theories of Heller and Nozick satisfy, is crucial. According to beta, the epistemic work done to know $\varphi \wedge \psi$ is guaranteed to be sufficient for one to know $\varphi \vee \psi$ only if the closest $(\neg\varphi \vee \neg\psi)$-worlds and the closest $(\neg\varphi \wedge \neg\psi)$-worlds overlap $(\mathsf{r}([\![\varphi \wedge \psi]\!], w) \cap \mathsf{r}([\![\varphi \vee \psi]\!], w) \neq \emptyset)$. It is unclear whether there is any intuitive motivation for this restriction of cover independent of the world-ordering picture.

Of the fallibilist theories we have considered so far, the $\mathsf{RS}_{\exists\forall}$ theories (Lewis, Sosa, DeRose) violate noVK and hence suffer from the Problem of Vacuous Knowledge, while the $\mathsf{RS}_{\forall\exists}$ theories (Nozick, Heller) violate cover and hence suffer from the Problem of Containment. In fact, the $\mathsf{RS}_{\exists\forall} + \mathsf{RO}_{\forall\exists}$ theories we have considered (Sosa, DeRose) suffer from both the Problem of Vacuous Knowledge, as observed in §4.1, and the Problem of Containment, as observed in Chapter 2 and highlighted in Fig. 4.4. The

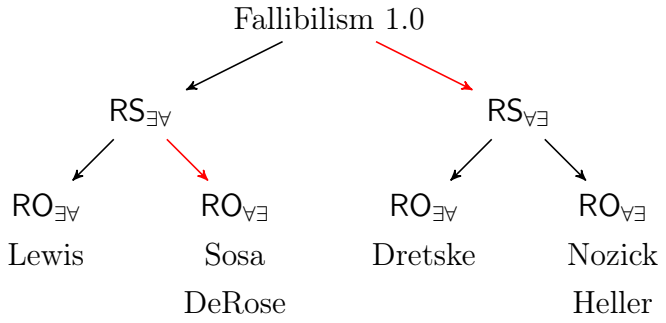

Figure 4.4: parameter settings and the Problem of Containment

reason is that although these theories satisfy cover for r, they do not satisfy the analogue of cover for u,

$$\textsf{u-cover} \quad \text{if } P \subseteq Q, \text{ then } \mathsf{u}(Q, w) \subseteq \mathsf{u}(P, w),$$

which says that if a possibility is uneliminated as an alternative for a proposition Q, then it is uneliminated as an alternative for any stronger proposition P. Recall the definition of the u function for safety and sensitivity theories from §3.3.2: v is uneliminated as an alternative for P by the agent in w, so $v \in \mathsf{u}(P, w)$, iff the agent falsely believes P in v. Since falsely believing a weaker proposition Q does not imply falsely believing a stronger P, the u-cover condition clearly fails. Now a symmetry between r and u becomes important: just as in Proposition 3.7, we showed that

> $K(\varphi \wedge \psi) \rightarrow K\varphi \wedge K\psi$ is valid on $\langle W, \mathsf{r}\rangle$ relative to models satisfying $\mathsf{RO}_{\exists\forall}$ iff r satisfies cover,

by a symmetrical argument we have

> $K(\varphi \wedge \psi) \rightarrow K\varphi \wedge K\psi$ is valid on $\langle W, \mathsf{u}\rangle$ relative to models satisfying $\mathsf{RS}_{\exists\forall}$ iff u satisfies u-cover.

Therefore, just as a $\mathsf{RO}_{\exists\forall}$ theory without cover fails to validate $K(\varphi \wedge \psi) \rightarrow K\varphi \wedge K\psi$, so does a $\mathsf{RS}_{\exists\forall}$ theory without u-cover fail to validate it. This explains why $K(\varphi \wedge \psi) \rightarrow K\varphi \wedge K\psi$ fails as a pure closure principle (recall Remark 2.1) for Sosa and DeRose.

These reflections on the important role of cover suggest a simple solution to the twin problems of Vacuous Knowledge and Containment. As theorists, we can postulate principles concerning what must be eliminated in order to know various propositions—we can postulate conditions on the r function—provided we have good reasons. Question: why not postulate that r satisfies both noVK and cover?

4.2.1 An Impossibility Result

Answer: because of the following simple impossibility result.

Proposition 4.1 (Impossibility I)**.** There is no SA model satisfying the following:

contrast	$\mathsf{r}(P, w) \subseteq \overline{P}$;
fallibilism	$\exists P \subseteq W$: $\overline{P} \not\subseteq \mathsf{r}(P, w)$;
noVK	if $P \neq W$, then $\mathsf{r}(P, w) \neq \emptyset$;
cover	if $P \subseteq Q$, then $\mathsf{r}(Q, w) \subseteq \mathsf{r}(P, w)$.

Proof. For *reductio*, suppose there is such a model. By fallibilism, there is some $P \subseteq W$ such that $\overline{P} \not\subseteq \mathsf{r}(P, w)$. Given contrast, it follows that

$$\mathsf{r}(P, w) \subsetneq \overline{P}. \tag{4.1}$$

Consider the proposition Q defined by

$$Q = P \cup \mathsf{r}(P, w). \tag{4.2}$$

Together (4.1) and (4.2) imply

$$Q \neq W. \tag{4.3}$$

Given $P \subseteq Q$, it follows by cover that $\mathsf{r}(Q, w) \subseteq \mathsf{r}(P, w)$, which with (4.2) implies

$$\mathsf{r}(Q, w) \subseteq Q. \tag{4.4}$$

However, contrast requires that $\mathsf{r}(Q, w) \subseteq \overline{Q}$, which with (4.4) implies

$$\mathsf{r}(Q, w) = \emptyset. \tag{4.5}$$

Together (4.3) and (4.5) contradict noVK. □

In Appendix §4.A I argue that moving to an alternatives-as-propositions picture does not avoid this kind of impossibility result, and in Appendix §4.B I argue that moving to more structured objects of knowledge does not help either.[10]

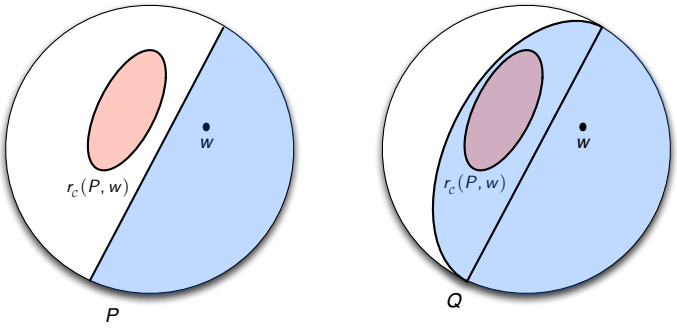

Figure 4.5: illustration for the proof of Proposition 4.1

Fig. 4.5 illustrates the key step in the proof of Proposition 4.1: where P is a contingent proposition that can be known without ruling out all not-P worlds, consider a contingent proposition Q that is true in all P-worlds *and* in all of the not-P worlds that one must rule out in order to know P. For example, Q could be the disjunction of P and various relevant counter-hypotheses that are true in all of the relevant not-P worlds; or Q could be the negation of a skeptical counter-hypothesis. In either case, it follows from contrast and cover that Q is a proposition that can be known without ruling out *any* possibilities, so we are back to vacuous knowledge.

Proposition 4.1 shows that we cannot avoid both the Problem of Vacuous Knowledge and the Problem of Containment in the framework of Fallibilism 1.0. Of course, there is one escape route that we have not yet considered: giving up the contrast condition and claiming that it is necessary in order to know Q that one rule out possibilities in which Q is true. But how could this be *necessary*? Surely ruling out *all* not-Q possibilities (even the most remote skeptical ones) should be enough to know Q. Suppose someone says, "I agree that you've ruled out every possible way in which

[10]Note added in ILLC version: see Holliday 2013b, §2.5 for a stronger version of the above impossibility result. Importantly, as noted in Holliday 2013b, §2.5, to derive a contradiction we do not need to assume that the conditions of Proposition 4.1 (e.g., noVK) hold for *all* propositions.

Q could be false, but that's not enough for you to know that Q is true; you also need to rule out such-and-such ways in which Q could be true." This seems absurd.

Or is it? Consider a Gettier case: not having any idea what time it is, you check a clock that—unbeknownst to you—has been stopped for weeks on 5:43; as it happens, the time is now 5:43; but you do not come to know this from the stopped clock. Where F is the proposition that *the time is 5:43* and S is the proposition that *the clock has stopped*, one might think this is a case in which knowing F *requires* ruling out $F \cap S$-possibilities, which would explain your ignorance of F (since you have not ruled those out) and violate contrast. But this is a mistake. What explains your ignorance of F is that since you have only looked at a stopped clocked, you have not ruled out various possibilities in which F is *false* and the time is something other than 5:43. If by some other means you had ruled out every possibility in which F is false, then it would be absurd to say "I agree you have ruled out every possibility in which the time is something other than 5:43, but you still do not know the time is 5:43 unless you rule out such-and-such possibilities in which the time is 5:43."

A similar point applies to *self-side* skeptical hypotheses. Imagine a skeptic who claims that in order to know that there is a Gadwall on the lake (P), not only must you rule out possibilities in which you are dreaming that there is a Gadwall on the lake when there is no Gadwall on the lake (let x be such an "unlucky dream" world), but also you must rule out possibilities in which you are dreaming that that there is a Gadwall on the lake *when there happens to be one* on the lake (let y be such a "lucky dream" world).[11] There are two different reasons one might think this:

1. The skeptic might think that *if* $y \in \mathsf{u}(P, w)$, *then* $x \in \mathsf{u}(P, w)$ (if you haven't ruled out the lucky dream world, then you haven't ruled out the unlucky dream world either), so given the skeptic's view that $x \in \mathsf{r}(P, w)$ (knowing P requires ruling out the unlucky dream world), knowing P requires $y \notin \mathsf{u}(P, w)$;

2. The skeptic might think that $y \in \mathsf{r}(P, w)$, so even if $\overline{P} \cap \mathsf{u}(P, w) = \emptyset$, knowing P still requires $y \notin \mathsf{u}(P, w)$.

While the first reason is intelligible, even skeptics should see the second as confused.

[11] Cf. Stroud 1984, 29.

If somehow you have ruled out every possible way that you could be wrong about their being a Gadwall on the lake, then you *know* that there is a Gadwall on the lake. There is no separate requirement that you rule out ways that you could be *right*.

I conclude that dropping contrast in the context of Fallibilism 1.0 is not an option. Therefore, by Proposition 4.1, the twin problems of Vacuous Knowledge and Containment seem to be inescapable for Fallibilism 1.0. In §4.3, I will consider a final attempt to escape, which leads to a third problem no less serious than the first two.

4.3 The Problem of Knowledge Inflation

In §4.1 I argued that $\mathsf{RS}_{\exists\forall}$ theories suffer from the Problem of Vacuous Knowledge, and in §4.2 I argued that $\mathsf{RS}_{\forall\exists}$ theories suffer from the Problem of Containment. In this section, I will consider an attempt to save $\mathsf{RS}_{\forall\exists}$ theories from the latter problem and sidestep Proposition 4.2.1, based on a defense of closure by Klein [1995].

Klein [1995, 216] begins by distinguishing two "sources of justification":

> We can conveniently divide the sources of justification into two mutually exclusive and jointly exhaustive types. One source is what I will call "externally situated evidence"—that is, features of the world other than the contents of S's actual beliefs and S's justified beliefs. During a murder investigation, the discovery of fingerprints, eyewitness testimony, letters, and traces of gunpowder may lead a detective to justifiably accuse someone of the crime. These are examples of *externally situated evidence*.
>
> On the other hand, the contents of a person's actual beliefs and justified beliefs can serve as an adequate source of justification for further beliefs. When the detective "puts two and two together" as, for example, when the detective recognizes the consequence of her belief that the murderer's fingerprints match those of a suspect, she may be led to justifiably believe the suspect is the murderer. Such potential sources of justification are *internally situated reasons*.

With this distinction, Klein offers a critique of the standard Dretske-style objections to closure. According to Klein [1995, 220]:

> [T]hese objections to the Closure Principle depend upon the fact that *the* externally situated evidence or *the* internally situated reasons that provide an adequate source of justification for a proposition, p, do not always provide an adequate source of justification for a proposition, q, entailed by p. But that fact cannot be used against the Closure Principle if the argument for closure depends upon the claim that in the relevant cases p, *itself, provides an adequate internally situated reason* for expanding the corpus of justified and/or known beliefs to those propositions obviously entailed by p. [last emphasis added]

Although here Klein discusses closure for justification rather than knowledge, we can consider an analogous defense of epistemic closure. To make Klein's point concrete, let us compare Klein's analysis with Stine's [1976] analysis (recall §4.1.1) applied to the Gadwall vs. Siberian Grebe example from Dretske [1981], our Example 1.2.

As fallibilists, Dretske, Stine, and Klein agree that (i) the birdwatcher's externally situated evidence e, obtained by observing the bird, is sufficient for him to know that *the bird is a Gadwall*, even though (ii) e does not rule out Siberian Grebe possibilities. Dretske concludes that the birdwatcher cannot know the denial of the Siberian Grebe hypothesis without more externally situated evidence, so closure fails.[12] In response, Stine attempts to save closure by saying that the birdwatcher can know the denial of the Siberian Grebe hypothesis on the basis of *no evidence*, because it is irrelevant. By contrast, Klein attempts to save closure by saying that although with only e the birdwatcher lacks sufficient externally situated evidence to know the denial of the Siberian Grebe hypothesis, the birdwatcher does have an *internally situated reason* r that is sufficient for her to know the denial of the Siberian Grebe hypothesis, namely that *the bird is a Gadwall*, which the birdwatcher knows on the basis of e.[13]

[12]Dretske does not actually discuss closure when he gives the Gadwall case in Dretske 1981, but this is the analysis he gives of the zebra case in Dretske 1971 (also see Dretske 2004, 2005).

[13]Here is Klein [1995, 221] on Dretske's [1971] zebra case:

> If we restrict the meaning of "evidence" or "reasons" (as used by Dretske) to what I

Cohen [1999] calls Klein's view *modus ponens fallibilism*[14] and remarks, "At first blush, it looks as if Klein has pulled the rabbit out of the hat" [2000, 101].[15] Brueckner [1998, 143] is on to the problem when he discusses Dretske's zebra case:[16]

> Klein's position can be seen in the following way. Proposition e is a good reason for believing z, z is a good reason for believing ∼cd, but e is not a good reason for believing ∼cd. The relation—is a good reason for believing . . . , then, is not transitive. . . . Note a peculiarity of this picture. Suppose that S does not infer from e to z and then to ∼cd, but rather reasons directly from e to ∼cd (as might an epistemologist who has run through the example a thousand times). Then S does not justifiably believe ∼cd, on Klein's view. This is because, according to the view we have reasonably attributed to Klein, e is not a good reason for believing ∼cd.

To say that this is a "peculiarity" is an understatement. I will argue that it leads to

---

> have called externally situated evidence, then Dretske is clearly correct. There can be adequate externally situated evidence to justify a proposition, p, without there being adequate externally situated evidence to justify a proposition, q, entailed by p. In addition, the internally situated reasons that are adequate for making p justified might not be adequate to make q justified. For example, the justified belief that the animals look like zebras and are in a pen marked "Zebras" cannot be used to justify the claim that the animals are not cleverly disguised mules. But the important point to note is that Dretske has restricted the search for a source of the justification of the entailed proposition in such a way that it precludes finding the entailing proposition—namely, the animals in the pen are zebras—as that source.

[14]Cohen [1999, 74] gives the following definition: "Let us say that when an alternative H, to P is eliminated on the basis of P, where the reasons for P are not reasons against H, that the reasons have an MPF [Modus Ponens Fallibilism] structure." Compare this with Brueckner's quote to follow.

[15]Cohen's [2000, 101] interpretation of Klein is essentially the same as mine:

> Klein agrees with Dretske that our evidence is sufficient for us to know that we see a Zebra. He also agrees that we cannot on the basis of our evidence come to know that we do not see a cleverly disguised mule—at least not directly. But on Klein's view, this poses no threat to the deductive closure principle. Since we can on the basis of our evidence come to know we see a Zebra, and since we can infer from the fact that we see a zebra, that we do not see a cleverly-disguised mule, we can thereby come to know that we do not see a cleverly-disguised mule.
>
> At first blush, it looks as if Klein has pulled the rabbit out of the hat.

[16]Brueckner takes 'z' to stand for the proposition that the animal is a zebra and '∼cd' for the proposition that the animal is not a cleverly disguised mule.

a serious Problem of Knowledge Inflation.[17] First, let us consider how to represent a view like Klein's in our framework, applying it to the Gadwall example.

Let P be the proposition that the bird is a Gadwall and S the proposition that the bird is a Siberian Grebe. According to Klein, if an agent goes directly from the observations of the bird to not-S, this will not suffice for knowledge of not-S; it will leave S-possibilities uneliminated that must be eliminated. However, if the agent goes from the same observations of the bird *first* to P and *then* to not-S, this will suffice for knowledge of not-S; all S-possibilities that must be eliminated will be eliminated.[18] What this suggests in our framework is a rule for updating the u function:

> Klein's Rule: if $\mathsf{r}(P,w) \cap \mathsf{u}(P,w) = \emptyset$ and the agent appropriately transitions from P to an entailed Q,[19] then update u to u' such that $\mathsf{r}(Q,w) \cap \mathsf{u}'(Q,w) = \emptyset$.[20]

With this rule, single-premise deductive closure holds even if cover fails for r, i.e., even if $\mathsf{r}(Q,w) \not\subseteq \mathsf{r}(P,w)$, and there is nothing stopping us from postulating noVK. Hence Klein's Rule promises to avoid the problems of Containment and Vacuous Knowledge. Unfortunately, by doing so it leads to the new Problem of Knowledge Inflation.[21]

We can illustrate the Problem of Knowledge Inflation by running a step-by-step analysis of the birdwatcher story. Before the birdwatcher has made any observations, let us ask: what will it take for the birdwatcher to come to know that bird b is a Gadwall? According to infallibilists, it will take eliminating *all* possibilities in which b is not a Gadwall, as shown by the red shading of the entire not-P zone on the right side of Fig. 4.6. By contrast, according to fallibilists, it will only take eliminating, e.g., possibilities in which b is of some other North American variety of bird, as shown on

[17] I am grateful to Helen Longino for proposing this apt name.

[18] As Cohen [1999, 75] puts it, "According to Klein, though sometimes an alternative to P must be "eliminated" prior to coming to know P, sometimes it can be eliminated after coming to know P, by appealing to P itself."

[19] Klein is not clear about the conditions under which an agent can come to know Q based on a newly acquired "internally situated reason" in the form of a newly known P that entails Q. Presumably the agent has to "put two and two together" or recognize the entailment, as in the quote from Klein about internally situated reasons, but he does not go into the details.

[20] Of course, this does not uniquely define an update rule, since there are many ways to update u to u' such that $\mathsf{r}(Q,w) \cap \mathsf{u}'(Q,w) = \emptyset$. However, for our purposes here it is enough to delimit a set of update rules with that effect. The criticism to follow will apply to any member of the set.

[21] In §5.4, I will propose a different rule for updating the u, which avoids this problem.

the left side of Fig. 4.6. Now let us consider three steps in the birdwatcher's inquiry:

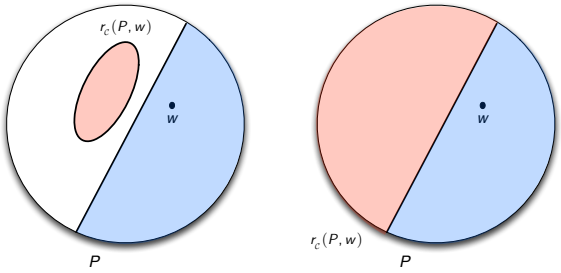

Figure 4.6: fallibilist picture (left) and infallibilist picture (right)

STEP I: Using her binoculars and guidebook, the birdwatcher eliminates possibilities in which b is of some other North American variety of bird, without eliminating skeptical possibilities in which b is a Siberian Grebe, animatronic robot, etc. See Fig. 4.7, where turning counter-possibilities from red to grey indicates their elimination.

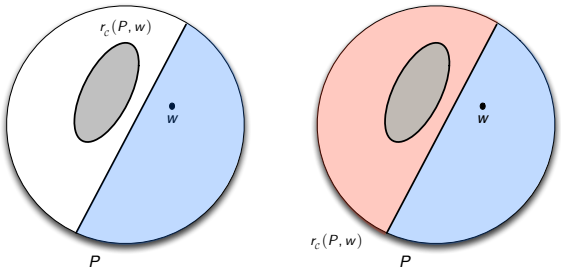

Figure 4.7: fallibilist picture (left) and infallibilist picture (right)

STEP II: According to fallibilists (including Klein), eliminating possibilities in which b is of some other North American variety of bird, without eliminating skeptical possibilities, is sufficient for the birdwatcher to know b is a Gadwall. See Fig. 4.8, where turning P from blue to green indicates the birdwatcher's new knowledge of P. Note that the infallibilist denies this new knowledge, so P is still blue on the right.

STEP III: According to Klein, although the birdwatcher came to know that b is a Gadwall by obtaining externally situated evidence that did not eliminate the skeptical

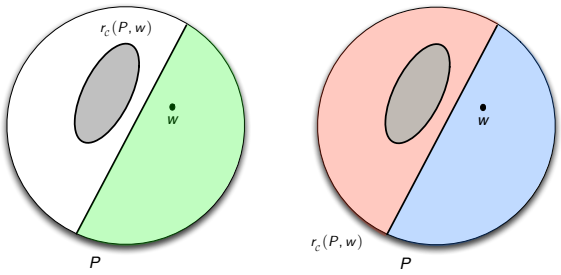

Figure 4.8: fallibilist picture (left) and infallibilist picture (right)

possibilities, having done so, she can turn right around and use that new internally situated reason—that b is a Gadwall—to eliminate *all* skeptical possibilities, thanks to closure. See Fig. 4.9, where the entire not-P zone has suddenly turned grey.

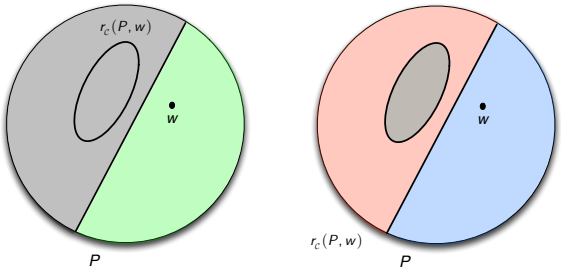

Figure 4.9: Klein's picture (left) and infallibilist picture (right)

It is in STEP III that *knowledge inflation* occurs. As Brueckner [1998, 146] says of Klein's view, "it is as if we are attempting to squeeze more out of S's evidence than is really there." Formally, there can be possibilities v such that before the new internally situated reason works its magic, v is uneliminated with respect to *all* propositions, including the P that gives the internally situated reason, but afterward (moving from u to u'), v becomes eliminated with respect to any Q entailed by P, including P itself:

$$\mathsf{KI} \quad v \in \mathsf{u}(P,w) \text{ for } \textit{all } P \subseteq W, \text{ but } v \notin \mathsf{u}'(Q,w).$$

To see this intuitively, consider the situation *just before* the birdwatcher crosses off the last North American variety she needs to eliminate: Pintail. At this point, Klein agrees that she cannot cross off skeptical possibilities. And yet Klein is committed to the view that as soon as she crosses off Pintail and comes to know that b is

a Gadwall, *then* she can cross off all of the skeptical possibilities using the new internally situated reason that b is a Gadwall. Moreover, the birdwatcher can do this even though her externally situated evidence that ruled out Pintail and the other North American varieties did not rule out the skeptical possibilities; if she had tried to use that evidence to rule out skeptical possibilities directly, she would have failed.

The problem with Klein's view is not with the idea of "internally situated reasons," as he first characterized them, but rather with the idea that these reasons are capable of inflating knowledge. The act of "putting two and two together" may extend knowledge, but it does not inflate knowledge as in Klein's picture. In §5.4, I will replace Klein's Rule for knowledge inflation with a new rule for knowledge extension.

Cohen [1999, 2000] also rejects Klein's view of "the structure of reasons," arguing that it licenses objectionable reasoning (see Cohen 1999, 74-76, Cohen 2000, 106, and Cohen 2002 on "easy knowledge"). Cohen [2000, 106] sums up the situation as follows:

> Our options seem to be accepting contingent *a priori* knowledge or endorsing what looks to be objectionable reasoning. However we go then, there is a distasteful consequence. But then again skepticism is a distasteful consequence—and I would maintain more so than any consequence of a contextualist account.
>
> I prefer . . . *a priori* rationality, but that may be mostly a statement about which bullet I am most prepared to bite.

In the framework of Fallibilism 1.0, it is true that however we go there is a distateful consequence and a bullet to bite: either the Problem of Vacuous Knowledge, the Problem of Knowledge Inflation, or the Problem of Containment. In my view, this is so much the worse for Fallibilism 1.0. While Cohen chooses to bite the first of these bullets, we will find a better way around them in Chapter 5: Fallibilism 2.0.

4.4 Conclusion

In this chapter, we have delved into the difficulties that arise from combining two attractive ideas: one idea is that the amount of epistemic work required to know

something can be limited (fallibilism); the other idea is that knowing something contingent always requires some amount of epistemic work (no vacuous knowledge). Unfortunately, when we put these ideas together, we seem forced to reject even weak closure (the containment problem) or accept that we can get more epistemic work out of inquiry than we put into it (knowledge inflation). The question is whether these difficulties are unavoidable—or whether they are artifacts of a flawed framework assumed by the standard fallibilist theories. The next chapter provides an answer.

4.A Alternatives as Possibilities vs. Propositions

In this section, I return to an issue raised in §3.1, the distinction between thinking of "alternatives" as possibilities/situations/scenarios/states of affairs or as more coarse-grained objects like propositions.[22] Both ways of thinking appear in the literature. For example, here is a brief historical survey of passages in the first tradition:

- "[L]et us consider a state of affairs in which it is true to say that it is possible, for all that the person ... knows, that p. Clearly the content of this statement cannot be adequately expressed by speaking of only one state of affairs. The statement in question can be true only if there is a possible state of affairs in which p would be true: but this state of affairs need not be identical with the one in which the statement was made. A description of such a state of affairs will be called an *alternative* .... (Sometimes the state of affairs will itself be said to be an alternative ....)" [Hintikka, 1962, 34].

- "A person knows that p, I suggest, only if the actual state of affairs in which p is true is distinguishable or discriminable by him from a relevant possible state of affairs in which p is false" [Goldman, 1976, 774].

[22]Note added in ILLC version: in Holliday 2013b, §2.1, I argue that if we take alternatives to be more coarse-grained than possibilities, then the set of alternatives (in a given context) should form a nontrivial *partition* of the set of possibilities. Views of the kind discussed in this appendix, according to which the set of alternatives for a proposition P is the set of *all propositions incompatible with* P, violate the partition condition because their "alternatives" are not mutually exclusive.

- "The social or pragmatic dimension to knowledge, if it exists at all, has to do with what counts as a relevant alternative, a possibility that must be evidentially excluded, in order to have knowledge" [Dretske, 1981, 367].

- "Which of all the uneliminated alternative possibilities may not properly be ignored? Which ones are the 'relevant alternatives'? - relevant, that is, to what the subject does and doesn't know" [Lewis, 1996, 554].

- "On RA, one need not be able to rule out every possibility of p's falsity in order to know p, but only the relevant alternatives to p" [Heller, 1999a].

- "According to premise 2, no experience gives an adequate basis for perceptual knowledge unless it enables the knower to discriminate situations where the believed proposition ⟨p⟩ is true, from all incompatible alternative situations. The "relevant alternatives" response is to deny that situations in which ⟨p⟩ is true must be discriminated by the knower from all ⟨p⟩-precluding alternatives. On the contrary, situations in which ⟨p⟩ is true must be discriminated only from those alternatives that are relevant" [Sosa, 2004, 35].[23]

On the other hand, here is a sampling of passages from the other tradition:

- "Let an alternative to a proposition q, be a proposition incompatible with q" [Cohen, 1988, 94].

- "An alternative A to a proposition P is a logical contrary of P; A is an alternative to P just in case P entails −A" [Vogel, 1999, 155].

- "As we'll understand the RA-Theory, it says that, if q is an irrelevant alternative to p, then knowing p doesn't require you to have evidence which would enable you to rule q out" [Pryor, 2001, 97].

Others, like Rysiew [2006], go back and forth between the two interpretations.

Let us set up our models and truth definition with alternatives-as-propositions.

[23]Sosa notes that "I do not distinguish formally between situations or scenarios or states of affairs that are actual and the propositions that fully capture such situations, etc., and are true. The reasoning of interest to us could be cast equivalently either way" (57).

Definition 4.1 (SAP Model)**.** A *standard alternatives-as-propositions* model is a tuple $\mathfrak{M}$ of the form $\langle W, \mathtt{u}, \mathtt{r}, V\rangle$ where W and V are as in Definition 3.1, $\mathtt{u}\colon \mathcal{P}(W) \times W \to \mathcal{P}(\mathcal{P}(W))$, and $\mathtt{r}\colon \mathcal{P}(W) \times W \to \mathcal{P}(\mathcal{P}(W))$.

Hence $\mathtt{r}$ sends each pair of a proposition P and a world w to a *set of propositions*, and similarly for $\mathtt{u}$. With this adjustment, the truth definition is as before.

Definition 4.2 (Truth in a SAP Model)**.** Given a SAP model $\mathfrak{M} = \langle W, \mathtt{u}, \mathtt{r}, V\rangle$ with $w \in W$ and a formula φ in the epistemic language, we define $\mathfrak{M}, w \vDash \varphi$ as follows (with propositional cases as usual):

$$\mathfrak{M}, w \vDash K\varphi \quad \text{iff} \quad \mathtt{r}(\llbracket\varphi\rrbracket^{\mathfrak{M}}, w) \cap \mathtt{u}(\llbracket\varphi\rrbracket^{\mathfrak{M}}, w) = \emptyset.$$

Clearly we can recover SA models as a special case of SAP models in which every set in $\mathtt{r}(P, w)$ and $\mathtt{u}(P, w)$ is a singleton set, so we have generalized.

As before, we can now consider constraints on the $\mathtt{r}$ and $\mathtt{u}$ functions, such as:

`r-constrast` $\mathtt{r}(P, w) \subseteq \mathcal{P}(\overline{P})$;

`u-constrast` $\mathtt{r}(P, w) \subseteq \mathcal{P}(\overline{P})$.

The study of SAP models can proceed from here using techniques similar to those used in the study of SA models. However, I will cut the study of SAP models short by arguing that viewing alternatives as propositions does not solve the fundamental problems with the framework of Fallibilism 1.0 discussed earlier in this chapter.

Let us begin with the question of whether *not-P* is in general a "relevant alternative" to P: is $\overline{P} \in \mathtt{r}(P, w)$? A positive answer threatens to plunge us into skepticism, for the following reason. I claim that any sensible theory of what it is to eliminate alternatives-as-propositions should satisfy the following condition:

`strength` if $Q \subseteq S$ and $Q \in \mathtt{u}(P, w)$, then $S \in \mathtt{u}(P, w)$.

This principle reflects the idea that eliminating a weaker proposition is at least as difficult as eliminating a stronger one. So, for example, if one has not eliminated the alternative *Mallard* to *Gadwall*, then one has not eliminated the weaker alternative $Mallard \vee Pintail$ to *Gadwall*. Whatever one's view of what it is to eliminate

alternatives-as-propositions, I think that one should accept the `strength` principle. One might argue that `strength` is like a closure condition, and if we are willing to give up some knowledge closure, then we should be willing to give up `strength`. However, this misunderstands the motivation for giving up some knowledge closure: the motivation has to do with how the range of alternatives—$\mathsf{r}(P, w)$—can be different for different propositions, which does not put any pressure on principles of elimination.

Returning to the plunge into skepticism, as an instance of `strength` we have

> if $Q \subseteq \overline{P}$ and $Q \in \mathtt{u}(P, w)$, then $\overline{P} \in \mathtt{u}(P, w)$.

Assuming `u-contrast`, $Q \in \mathtt{u}(P, w)$ implies $Q \subseteq \overline{P}$, so the instance reduces to

> if $Q \in \mathtt{u}(P, w)$, then $\overline{P} \in \mathtt{u}(P, w)$.

Now if $\overline{P} \in \mathtt{r}(P, w)$, we have the following result:

> if $Q \in \mathtt{u}(P, w)$, then $\mathtt{r}(P, w) \cap \mathtt{u}(P, w) \neq \emptyset$;

so if there is *any* uneliminated alternative to P, then the agent does not know P, which amounts to skepticism given the inevitability of uneliminated alternatives.

I conclude from this that $\overline{P}$ cannot in general count as a relevant alternative to P in this setting, if we are to maintain fallibilism. Hence it is not enough to define fallibilism by analogy with our definition for alternatives-as-possibilities[24] as

> `fallibilism`$^-$ $\exists P \subseteq W$: $\mathcal{P}(\overline{P}) \not\subseteq \mathtt{r}(P, w)$,

since this is compatible with $\overline{P} \in \mathtt{r}(P, w)$. Instead, we must *at least* require

> `fallibilism` $\exists P \subseteq W$: $\overline{P} \notin \mathtt{r}(P, w)$.

Whether this captures enough of fallibilism is a further question. However, we are already on the path to an impossibility result analogous to the result of §4.2.1. As the final step along this path, I claim that any sensible theory of the relevance of alternatives-as-propositions should satisfy the principle that a *disjunction* of relevant alternatives is itself a relevant alternative:

[24] Recall the condition of fallibilism: $\exists P \subseteq W$: $\overline{P} \not\subseteq \mathsf{r}(P, w)$.

`r-union` if $\Sigma \subseteq \mathtt{r}(P, w)$, then $\bigcup \Sigma \in \mathtt{r}(P, w)$.

For example, if *Mallard* is a relevant alternative to *Gadwall*, and *Pintail* is a relevant alternative to *Gadwall*, then it hardly makes sense to claim that $Mallard \vee Pintail$ is an irrelevant alternative to *Gadwall*.

We can now see that the move to alternatives-as-proposition does not solve the problems of §4, given Proposition 4.2. Let us define:

`noVK` if $P \neq W$, then $\mathtt{r}(P, w) \neq \emptyset$;
`cover` if $P \subseteq Q$, then $\mathtt{r}(Q, w) \subseteq \mathtt{r}(P, w)$.

The `noVK` and `cover` conditions have the same form and meaning in the alternatives-as-propositions setting as noVK and cover did in the alternatives-as-possibilities setting: `noVK` says that knowing a contingent proposition requires epistemic work in the sense of ruling out some alternative(s), while `cover` says that if P excludes as much of logical space as Q does, then coming to know P requires at least as much epistemic work, in terms of ruling out alternatives, as coming to know Q does.

4.A.1 Another Impossibility Result

Unfortunately, the principles discussed so far cannot be consistently combined.

Proposition 4.2 (Impossibility II)**.** There is no SAP model satisfying the following conditions:

`r-contrast` $\mathtt{r}(P, w) \subseteq \mathcal{P}(\overline{P})$;
`fallibilism` $\exists P \subseteq W$: $\overline{P} \notin \mathtt{r}(P, w)$;
`noVK` if $P \neq W$, then $\mathtt{r}(P, w) \neq \emptyset$;
`cover` if $P \subseteq Q$, then $\mathtt{r}(Q, w) \subseteq \mathtt{r}(P, w)$;
`r-union` if $\Sigma \subseteq \mathtt{r}(P, w)$, then $\bigcup \Sigma \in \mathtt{r}(P, w)$.

Proof. For *reductio*, suppose there is such a model. By `fallibilism`, there is some $P \subseteq W$ such that $\overline{P} \notin \mathtt{r}(P, w)$. Given `r-union`, it follows that

$$\bigcup \mathtt{r}(P, w) \neq \overline{P}, \tag{4.6}$$

so given `r-contrast` it follows that

$$\bigcup \mathtt{r}(P, w) \subsetneq \overline{P}. \tag{4.7}$$

Consider the proposition Q defined by

$$Q = P \cup \bigcup \mathtt{r}(P, w). \tag{4.8}$$

Together (4.7) and (4.8) imply

$$Q \neq W. \tag{4.9}$$

Given $P \subseteq Q$, it follows by `cover` that $\mathtt{r}(Q, w) \subseteq \mathtt{r}(P, w)$, which with (4.8) implies

$$\mathtt{r}(Q, w) \subseteq \mathcal{P}(Q). \tag{4.10}$$

However, `r-contrast` requires that $\mathtt{r}(Q, w) \subseteq \mathcal{P}(\overline{Q})$, which with (4.10) implies

$$\mathtt{r}(Q, w) = \emptyset. \tag{4.11}$$

Together (4.9) and (4.11) contradict `noVK`. □

I conclude that the move to alternatives-as-propositions does not fix the flaws of Fallibilism 1.0. To avoid impossibility results like Propositions 4.2 and 4.1, we must depart from Fallibilism 1.0 in more fundamental ways, as in Chapter 5.

4.B Structured Objects of Knowledge

So far we have assumed that the objects of knowledge are no more finely individuated than sets of worlds. In this section, I show that even if we assume that the objects of knowledge are as finely individuated as formulas in our language, this additional structure will not solve the problems for Fallibilism 1.0 raised in §4.2.1. To show this, we define a new class of models as follows. Form is the set of formulas in our language.

Definition 4.3 (SSA Model)**.** A *fine-grained standard alternatives* model is a tuple M of the form $\langle W, \mathrm{u}, \mathrm{r}, V\rangle$ where W and V are as in Definition 3.1, $\mathrm{u}\colon \mathsf{Form}\times W \to \mathcal{P}(W)$, and $\mathrm{r}\colon \mathsf{Form} \times W \to \mathcal{P}(W)$.

Here $\mathrm{r}(\varphi, w)$ is the set of possibilities that the agent must eliminate to know φ in world w, and $\mathrm{u}(\varphi, w)$ is the set of possibilities that the agent has not eliminated as alternatives for φ in w. The truth definition is analogous to that of Definition 3.2.

Definition 4.4 (Truth in a SSA Model)**.** Given a SSA model $M = \langle W, \mathrm{u}, \mathrm{r}, V\rangle$ with $w \in W$ and a formula φ in the epistemic language, we define $\mathfrak{M}, w \vDash \varphi$ as follows (with propositional cases as usual):

$$M, w \vDash K\varphi \quad \text{iff} \quad \mathrm{r}(\varphi, w) \cap \mathrm{u}(\varphi, w) = \emptyset.$$

With this setup, conditions on SA models such as

$\mathsf{contrast}$	$\mathsf{r}(P, w) \subseteq \overline{P}$,
noVK	if $P \neq W$, then $\mathsf{r}(P, w) \neq \emptyset$, and
$\mathsf{RO}_{\exists\forall}$	$\exists \mathsf{U}(w) \subseteq W\ \forall P \subseteq W\colon \mathsf{u}(P, w) = \mathsf{U}(w) \cap \overline{P}$,

have analogues on SSA models such as

$\mathsf{contrast}$	$\mathsf{r}(\varphi, w) \subseteq \overline{[\![\varphi]\!]}$,
noVK	if $[\![\varphi]\!] \neq W$, then $\mathrm{r}(\varphi, w) \neq \emptyset$, and
$\mathrm{RO}_{\exists\forall}$	$\exists \mathsf{U}(w) \subseteq W\ \forall \varphi\colon \mathsf{u}(\varphi, w) = \mathsf{U}(w) \cap \overline{[\![\varphi]\!]}$.

In addition, we can state conditions on SSA models that correspond to closure properties in a more fine-grained manner than before. For example, the principle $K\varphi \to K(\varphi \vee \psi)$ corresponds (relative to models satisfying $\mathrm{RO}_{\exists\forall}$) to

$\vee$-cover	$\mathrm{r}(\varphi \vee \psi, w) \subseteq \mathrm{r}(\varphi, w)$.

Similarly for $K(\varphi \wedge \psi) \to (K\varphi \wedge K\psi)$:

$\wedge$-cover	$\mathrm{r}(\varphi, w) \cup \mathrm{r}(\psi, w) \subseteq \mathrm{r}(\varphi \wedge \psi, w)$.

Hence we can consider classes of SSA models that validate one of these closure principles but not the other. Finally, let us consider how to express fallibilism in SSA models. By analogy with $\mathsf{fallibilism}$ in SA models, we could define fallibilism in SSA

models as follows: there is some φ such that $\overline{[\![\varphi]\!]} \not\subseteq \mathrm{r}(\varphi, w)$. However, I will use a slightly stronger definition: there are some φ and ψ such that $\mathrm{r}(\varphi, w) \subseteq [\![\psi]\!]$ and $\overline{[\![\varphi]\!]} \not\subseteq [\![\psi]\!]$. Intuitively, this says that there is some ψ that is true in all of "relevant" $\neg\varphi$-worlds (perhaps in exactly these worlds) but *not* true in all of the "irrelevant" $\neg\varphi$-worlds. In other words, some difference between the set of relevant $\neg\varphi$-worlds and the set of irrelevant $\neg\varphi$-worlds is expressible in our language.[25]

Despite the additional flexibility of SSA models, §4.B.1 shows that they are not flexible enough to avoid both the problems of Vacuous Knowledge and Containment.

4.B.1 More Impossibility Results

The following proposition gives one of many analogues to Proposition 4.1.

Proposition 4.3 (Impossibility IB)**.** There is no SSA model satisfying the following:

contrast	$\mathrm{r}(\varphi, w) \subseteq \overline{[\![\varphi]\!]}$;
fallibilism	$\exists\varphi, \psi$: $\mathrm{r}(\varphi, w) \subseteq [\![\psi]\!]$ and $\overline{[\![\varphi]\!]} \not\subseteq [\![\psi]\!]$;
noVK	if $[\![\varphi]\!] \neq W$, then $\mathrm{r}(\varphi, w) \neq \emptyset$;
$\vee$-cover	$\mathrm{r}(\varphi \vee \psi, w) \subseteq \mathrm{r}(\varphi, w)$.

Proof. For *reductio*, suppose there is such a model. By fallibilism, there are some φ and ψ such that

$$\mathrm{r}(\varphi, w) \subseteq [\![\psi]\!] \tag{4.12}$$

and

$$\overline{[\![\varphi]\!]} \not\subseteq [\![\psi]\!]. \tag{4.13}$$

Consider $\varphi \vee \psi$. By (4.13), we have

$$[\![\varphi \vee \psi]\!] \neq W. \tag{4.14}$$

By $\vee$-cover, we have $\mathsf{r}(\varphi \vee \psi, w) \subseteq \mathsf{r}(\varphi, w)$, which with (4.12) implies

$$\mathsf{r}(\varphi \vee \psi, w) \subseteq [\![\varphi \vee \psi]\!]. \tag{4.15}$$

[25]Note added in ILLC version: see the discussion of "expressible fallibilism" in Holliday 2013b, §2.1, and compare Propositions 4.3 - 4.4 below to Proposition 1 in Holliday 2013b, §2.5.

However, contrast requires that $\mathsf{r}(\varphi \vee \psi, w) \subseteq \overline{[\![\varphi \vee \psi]\!]}$, which with (4.15) implies

$$\mathsf{r}(\varphi \vee \psi, w) = \emptyset. \tag{4.16}$$

Together (4.14) and (4.16) contradict noVK. □

Proposition 4.3 shows that even with SSA models, we cannot avoid the Problem of Vacuous Knowledge while also validating the principle $K\varphi \to K(\varphi \vee \psi)$. The following corollary shows that we cannot avoid the Problem of Vacuous Knowledge while also validating the principles $K(\varphi \wedge \psi) \to K\varphi$ and $K\varphi \leftrightarrow K((\varphi \vee \psi) \wedge \varphi)$, the latter being a special case of closure under logical equivalence.

Proposition 4.4 (Impossibility IC)**.** There is no SSA model satisfying the following:

contrast	$\mathrm{r}(\varphi, w) \subseteq \overline{[\![\varphi]\!]}$;
fallibilism	$\exists \varphi, \psi$: $\mathrm{r}(\varphi, w) \subseteq [\![\psi]\!]$ and $\overline{[\![\varphi]\!]} \not\subseteq [\![\psi]\!]$;
noVK	if $[\![\varphi]\!] \neq W$, then $\mathrm{r}(\varphi, w) \neq \emptyset$;
$\wedge$-cover	$\mathrm{r}(\varphi, w) \subseteq \mathrm{r}(\varphi \wedge \psi, w)$;
$\vee\wedge$-equiv	$\mathrm{r}(\varphi, w) = \mathrm{r}((\varphi \vee \psi) \wedge \varphi, w)$.

Proof. By $\wedge$-cover and $\vee\wedge$-equiv, we have

$$\mathrm{r}(\varphi \vee \psi, w) \subseteq \mathrm{r}((\varphi \vee \psi) \wedge \varphi, w) = \mathrm{r}(\varphi, w),$$

so $\vee$-cover holds. Hence there is no such model by Proposition 4.3. □

Note that $\mathrm{r}(\varphi, w) = \mathrm{r}((\varphi \vee \psi) \wedge \varphi, w)$ says that for an IAL, coming to know φ requires empirically eliminating the same alternatives as coming to know the logically equivalent $(\varphi \vee \psi) \wedge \varphi$. Assuming this modest principle, Proposition 4.4 shows that in Fallibilism 1.0, we must either accept vacuous knowledge or give up $K(\varphi \wedge \psi) \to K\varphi$.

I conclude that the move to more structured objects of knowledge does not fix the flaws of Fallibilism 1.0. To avoid impossibility results like Propositions 4.3 and 4.4, we must depart from Fallibilism 1.0 in more fundamental ways, as in Chapter 5.

5

Fallibilism 2.0: The Multipath Picture

In Chapter 3, we studied how the framework of Fallibilism 1.0 unifies the RA and subjunctivist theories from Chapter 2. This unification revealed in Chapter 4 how problems with those theories are manifestations of a trio of problems facing any theory developed in the same framework: the Problem of Vacuous Knowledge, the Problem of Containment, and the Problem of Knowledge Inflation. More tinkering within the framework of Fallibilism 1.0 will not solve these problems. What is required is a more fundamental change. As promised in Chapter 4, we will now apply what we have learned about the flaws of Fallibilism 1.0 in search of a new and improved framework of Fallibilism 2.0, with the goal of resolving the three problems.

In this chapter, I offer a proposal for Fallibilism 2.0 in the form of what I call the Multipath Picture of Knowledge. This picture of knowledge is based on taking seriously the idea that there can be multiple paths to knowing a complex claim. An overlooked consequence of fallibilism is that these multiple paths to knowledge of a claim may involve ruling out different sets of alternatives, which should be represented in our picture of knowledge. I will argue that the Multipath Picture of Knowledge is a better picture for all fallibilists, whether for or against closure, whether for or against contextualism, compared to the "single path picture" assumed by Fallibilism 1.0. I will also argue for giving up full closure, but those who accept full closure will

do better with the Multipath Picture than without it. Finally, I will present a new picture of the epistemic effect of "putting two and two together," the Transfer Picture of Deduction, showing how deduction can extend knowledge without inflating it.

5.1 Back to the Drawing Board

Recall the starting point of Fallibilism 1.0 from Chapter 3: for each proposition to be known, there is "*a set* of situations each member of which *contrasts* with what is [to be] known...and must be evidentially excluded if one is to know" [emphasis added] [Dretske, 1981, 373]. Against these *contrast* and *set* assumptions, I will argue:

- In some cases, it is sufficient (as far as empirical work goes) for an agent to know P that she *only rules out non-contrasting possibilities in which P is* **true**.
- In some cases, there is no set of situations all of which must be excluded if one is to know; instead, there are **multiple sets** of situations, such that if one is to know, one must exclude *all of the situations in at least one of those sets.*

In the following subsections, I will argue for both of these claims in turn.

5.1.1 Against the Single Alternative Set Assumption: The Multipath Picture of Knowledge

Suppose an agent wants to know whether $\alpha \vee \beta$ is true, where α and β are contingent. Further suppose that it is true. Then there are at least three paths by which she could come to know it: she could start with α, and if she comes to know α, then she is done (at least with ruling out possibilities); or she could start with β, and if she comes to know β, then she is done (with ruling out possibilities); or she could come to know that $\alpha \vee \beta$ is true without coming to know which disjunct is true[1] (by, e.g.,

[1]Examples of coming to know along the third path abound. I can know by observation that either horse A won the race or horse B won the race, since they were far ahead of all of the other horses, even though I do not know that horse A won the race and I do not know that horse B won the race, since the finish was too close for me to tell. Testimony also provides many examples.

ruling out all $(\neg\alpha \wedge \neg\beta)$-possibilities without ruling out any $(\neg\alpha \wedge \beta)$-possibilities or any $(\alpha \wedge \neg\beta)$-possibilities). This is just common sense. But it raises the question of why anyone should think that for something like $\alpha \vee \beta$, there is a *single* set of situations that must be evidentially excluded if one is to know $\alpha \vee \beta$. It seems instead that there should be at least *three* sets of situations such that if one is to know $\alpha \vee \beta$, one must evidentially exclude all of the situations in at least one of those three sets, corresponding to the three paths to knowledge of $\alpha \vee \beta$ described above.

If we were *infallibilists*, then there would be no need for these multiple "alternative sets" for $\alpha \vee \beta$. According to infallibilism, coming to know α requires (among other things) ruling out all $(\neg\alpha \wedge \neg\beta)$-possibilities; so does coming to know β; and so does coming to know $\alpha \vee \beta$ without coming to know which disjunct is true. Moreover, as argued in §4.2.1, ruling out all $\neg\varphi$-possibilities should be *sufficient* for knowing φ. Therefore, infallibilists need only consider one alternative set for $\alpha \vee \beta$: to know $\alpha \vee \beta$ it is necessary and sufficient that one rule out all $(\neg\alpha \wedge \neg\beta)$-possibilities.

But we are *fallibilists*. According to fallibilism, coming to know α may not require ruling out *all* $(\neg\alpha \wedge \neg\beta)$-possibilities. Indeed, it may not require ruling out *any* $(\neg\alpha \wedge \neg\beta)$-possibilities,[2] e.g., if $\neg\beta$ is a skeptical hypothesis.[3] But then since it suffices to know $\alpha \vee \beta$ that one rule out all $(\neg\alpha \wedge \neg\beta)$-possibilities, it is immediate that we need multiple alternative sets for $\alpha \vee \beta$, corresponding to the multiple paths to knowledge of $\alpha \vee \beta$ described above: the $(\neg\alpha \wedge \neg\beta)$-possibilities that one must rule

[2]Note added in ILLC version: indeed, the claim that for every α and β, knowing α requires ruling out some $(\neg\alpha \wedge \neg\beta)$-possibility (if there is one) is essentially equivalent to infallibilism; and if every possibility were definable, then the claim would be exactly equivalent to infallibilism. Given any $\neg\alpha$-possibility v, pick β so that $\neg\beta$ is true only at v. Then if knowing α requires ruling out some $(\neg\alpha \wedge \neg\beta)$-possibility (if there is one), the agent must rule out v. Since v was arbitrary, it follows that knowing α requires ruling out every $\neg\alpha$-possibility, which is the infallibilist position.

[3] Contextualists should read this as a claim about what the agent must rule out in order to know $\alpha \vee \beta$ relative to a context in which skeptical possibilities are irrelevant. A Lewis-style contextualist (recall §4.1.2) may claim that if $\neg\beta$ is a skeptical hypothesis, then the mere mention of β will shift the context to one in which knowing α requires ruling out all $(\neg\alpha \wedge \neg\beta)$-possibilities. Setting aside the problems with this view discussed in §4.1.2, it is still compatible with everything I say here. We simply observe Hawthorne's point from §4.1.2: even if the *agent* is considering a skeptical hypothesis $\neg\beta$ at time t (let us assume she is fully confident in $\alpha \vee \beta$), if the *attributors* are *not* considering any skeptical hypotheses, then they can truly say "the agent knows α at t, so she knows any disjunction $\alpha \vee X$ that she has derived from α at t," which requires for its truth that the agent can know α relative to the attributor's context despite not having ruled out skeptical $(\neg\alpha \wedge \neg\beta)$-possibilities.

out in order to know one disjunct may be different from those that one must rule out in order to know the other disjunct, which may be different from those that one must rule out in order to know the disjunction without knowing either disjunct.

What this shows is that we should replace the r function of Fallibilism 1.0, which assigns to each pair of a proposition P and a world w a set $\mathsf{r}(P, w) \subseteq W$ of possibilities, with a new $\mathbf{r}$ function for Fallibilism 2.0 that assigns to each pair of a proposition P and a world w a set $\mathbf{r}(P, w) = \{A_1, A_2, \dots\}$ *of sets* $A_i \subseteq W$ of possibilities. For example, for $\alpha \vee \beta$ we may have alternative sets A_1, A_2, and A_3, where A_1 is the set of possibilities to be ruled out in the path to knowing $\alpha \vee \beta$ that goes via α; A_2 is the set of possibilities to be ruled out in the path that goes via β; and A_3 is the set of possibilities to be ruled out in the path to knowing $\alpha \vee \beta$ without knowing either α or β. (I am intentionally sliding between the proposition P and the formula $\alpha \vee \beta$ until §5.1.4.) Later in the chapter we will see concrete examples of this form.

Although here we are working with a propositional language, the foregoing points about disjunctive claims clearly apply to existential claims as well. One could come to know $\exists x \varphi(x)$ by coming to know $\varphi(a)$, or by coming to $\varphi(b)$, etc., or by coming to know $\exists x \varphi(x)$ without coming to know $\varphi(c)$ for any c. As a consequence of fallibilism, the alternatives sets for these different paths to knowing $\exists x \varphi(x)$ may be different.

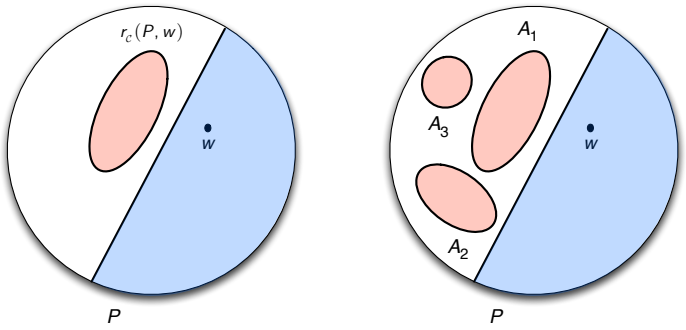

Figure 5.1: Single-Path Picture (left) vs. Multipath Picture (right)

I have used the term 'path to knowledge' instead of 'way of knowing'. There are often "multiple ways of knowing" a claim in the sense that one can come to know the claim by ruling out a *single set* of alternatives in a number of ways: by sight, sound, smell, etc. I reserve the idea of "multiple paths to knowledge" for the case in which

for a given claim there are *multiple sets of alternatives* such that in order to know the claim, it suffices to rule out all of the alternatives in one of those sets (which one may often do in a number of *ways*). The multiplicity of paths arises from the structure of the claim itself, rather than variation in the methods of inquiry. Indeed, it arises from the *logical* structure of the claim, to which I return in §5.1.4 and §5.2.4.

The move from r to $\mathbf{r}$ is the basis of what I call the Multipath Picture of Knowledge. In §5.2.1, I will propose Five Postulates for properties of $\mathbf{r}$ in the Multipath Picture. First, however, I must explain why one of the fundamental assumptions of Fallibilism 1.0 must be dropped when we adopt the Multipath Picture for Fallibilism 2.0.

5.1.2 Against the Contrast Assumption

Recall the contrast assumption from Fallibilism 1.0:

$$\textsf{contrast} \quad \mathsf{r}(P, w) \subseteq \overline{P}.$$

In §4.2.1, I argued that we cannot give up contrast in the framework of Fallibilism 1.0 (to avoid the impossibility result of Proposition 4.1), because it should always be sufficient for knowing a true proposition P that one rule out *all* not-P possibilities.

In the Multipath Picture of Knowledge, the corresponding assumption is

$$\textbf{contrast} \quad \bigcup \mathbf{r}(P, w) \subseteq \overline{P},$$

which says that all alternative sets for P are sets of not-P possibilities. In other words, the assumption is that all paths to knowing P only involve eliminating not-P possibilities. Should fallibilists accept this assumption? The answer is 'no', as we have already seen a counterexample in §5.1.1. If one path to knowing $\alpha \vee \beta$ is via knowing α, and if knowing α only requires ruling out $(\neg\alpha \wedge \beta)$-possibilities (e.g., because $\neg\beta$ is a skeptical hypothesis), which are of course $(\alpha \vee \beta)$-possibilities, then there is a path to knowing $\alpha \vee \beta$ that only involves ruling out $(\alpha \vee \beta)$-possibilities. (This is so even for the shiftiest versions of contextualism, for reasons explained in footnote 3.) Since the idea that knowing α is a path to knowing $\alpha \vee \beta$ is at the core

of common sense about knowledge, and the idea that knowing α may only require ruling out $(\neg\alpha \wedge \beta)$-possibilities is at the core of fallibilism, we must reject **contrast**.

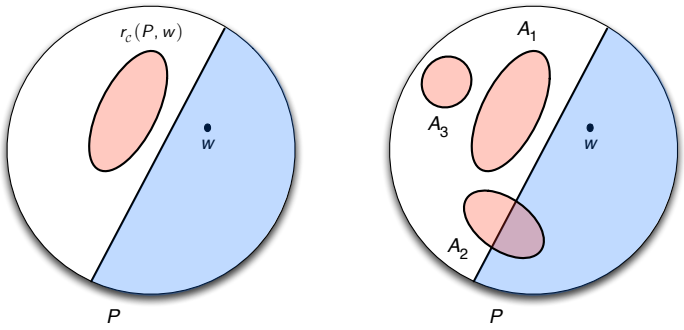

Figure 5.2: Single-Path Picture with contrast (left) vs. Multipath Picture without **contrast** (right)

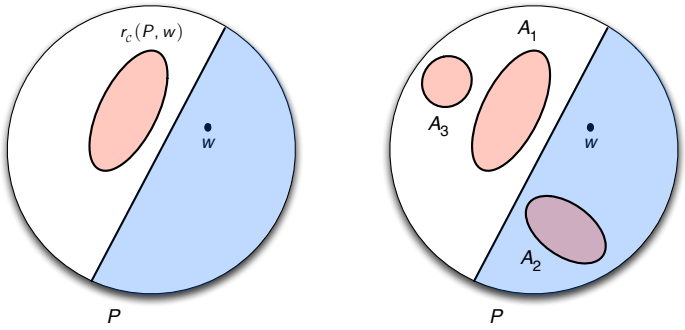

Figure 5.3: Single-Path Picture with contrast (left) vs. Multipath Picture without **contrast** (right)

Having just argued for the strong claim that there can be alternatives sets $A \in \mathbf{r}(P, w)$ such that $A \subseteq P$, we also have the weaker claim that there can be alternative sets $A \in \mathbf{r}(P, w)$ such that $A \cap P \neq \emptyset$. Later I will propose an **overlap** postulate concerning the conditions under which an alternative set for P may overlap with P in this way. For now, however, let us observe that in the Multipath Picture, denying **contrast** is compatible with holding that it should always be sufficient for knowing

a true proposition P that one rule out all not-P possibilities, guaranteed as follows:[4]

$$\textbf{enough} \quad \text{if } w \in P, \text{ then } \exists A \in \mathbf{r}(P, w) \colon A \subseteq \overline{P}.$$

(Applying this to our $\alpha \vee \beta$ example, there should be an alternative set A for $\alpha \vee \beta$ such that A is included in the set of $\neg(\alpha \vee \beta)$-worlds, representing the path to knowing $\alpha \vee \beta$ without necessarily knowing either disjunct, in addition to the alternative set for the α-path that overlaps with the set of $(\alpha \vee \beta)$-worlds.) What this shows is that the argument for the necessity of contrast in the Single-Path Picture of Fallibilism 1.0, given in §4.2.1, does not apply to **contrast** in the Multipath Picture.

It is worth emphasizing that the foregoing arguments for the Multipath Picture without **contrast** should be acceptable to fallibilists who support full closure as much as to those who do not. As we will see, the Multipath Picture without **contrast** will be key to resolving the trio of problems for Fallibilism 1.0 discussed in Chapter 4. Before resolving those problems, however, we must discuss one more aspect of our basic framework: in Figs. 5.1 - 5.3, what kind of space do the large circles represent?

5.1.3 Logical Space

In the previous chapters, I left open how exactly we should think of the domain W of our RA models (§2.4), CB models (§2.5), and SA models (Chapter 3). Here I will be explicit about how we should think of W for the purposes of this chapter.

My approach to this issue is not metaphysical, in the style of Lewis [1986], but rather pragmatic, in the style of Stalnaker [1984, 57]:

> To believe in possible worlds is to believe only that [rational] activities have a certain structure, the structure which possible worlds theory helps to bring out.

It is important to realize that those who agree with Stalnaker that to "believe" in a possible worlds picture is just to believe that rational activities have the structure that

[4]Note added in ILLC version: a simpler version of **enough** simply states that for all propositions P, $\exists A \in \mathbf{r}(P, w) \colon A \subseteq \overline{P}$. Using this simpler version requires changing the **r-RofA** condition in Definition 5.3 from $w \in \overline{P} \Rightarrow \mathbf{r}(P, w) = \emptyset$ to $w \in \overline{P} \Rightarrow w \in \bigcap \mathbf{r}(P, w)$, as in Holliday 2013b.

the picture helps bring out may nonetheless disagree with him about what should be included in the picture, because they disagree with him about matters of structure.

In Stalnaker's picture (as in Lewis's), W contains only "metaphysically possible worlds," sets of which are propositions. As a result, there is no way to capture in a model the distinction between *metaphysically equivalent* but *logically inequivalent* propositions. For if two claims are logically inequivalent but are true in the same metaphysically possible worlds, then they express the same proposition for Stalnaker. In my view, this a serious problem for modeling knowledge in a Stalnakerian picture, for at least two reasons: one general and one specific to fallibilist views of knowledge. (Moreover, these problems are distinct from the kind of worries that arise when we treat logically equivalent claims as expressing the same proposition.) I will explain the general problem here and the specific problem for fallibilism in §5.3.

First, however, let us note two modifications that would allow us to distinguish propositions that are metaphysically equivalent but logically inequivalent. One option is to stick with propositions as sets, but allow more into these sets than only metaphysically possible worlds. Another option is to replace propositions as sets by some more fine-grained objects or structured propositions (see §5.1.4 and §5.2.4).

If we stick with sets, then the general problem with restricting W to include only "metaphysically possible worlds" is clearly stated by Kaplan [1995], who asks what would happen to possible worlds semantics if metaphysicians discovered the truth of Hyperdeterminism, the thesis that the only metaphysically possible world is the actual world. Kaplan's [1995, 48] discussion is worth quoting at length:

> The metaphysical should not be confused with the logical. If Hyperdeterminism were to imply that there is only one true proposition, then not only would whatever is true be necessary (as would be expected), but *all* propositional operators would become truth functional. Even if this were the only metaphysically possible world, should 'it is desirable that' and 'it is undesirable that' become truth functional? I think not!
>
> A proper PWS framework for a language containing both *possibility* and *desirability* operators, should I believe, allow the logical to dominate the

> metaphysical .... This means that Hyperdeterminism or not, we must retain all the points (representing so-called possible worlds) needed to distinguish the propositions expressed by [logically] inequivalent sentences .... It would be reasonable to take the view that real possible worlds correspond to some of the points, namely, those that are metaphysically possible.

Note that Kaplan's remarks apply as much to 'it is known that' as they do to 'it is desirable that': even if the actual world were the only metaphysically possible world, 'it is known that' should not become truth functional, so a proper PWS framework for a language containing both *possibility* and *knowledge* operators should allow the logical to dominate the metaphysical in Kaplan's sense.[5] (Of course, practitioners of epistemic logic and modal logic more generally follow Kaplan and not Stalnaker on this point: they include in their models—especially canonical models—whatever points, with whatever propositional valuations, are needed to make logical distinctions.)

Salmon [1989] presents metaphysical arguments for admitting *total ways things could not have been* ("metaphysically impossible worlds") as well as *total ways things could have been* ("metaphysically possible worlds") and for considering metaphysical necessity as a restriction of logical necessity. King [2007] argues that Stalnaker's pragmatic methodological approach to possible worlds supports the inclusion of *ways things count not have been* as well. However, I will not go into these accounts here. In our formal framework, Kaplan's extra points are unmysterious: they are points whose associated propositional valuations do not correspond to a metaphysically possible world according to the intended meaning of the sentence letters $p, q, \ldots$.[6]

My unmetaphysical attitude about the extra points is justified by the weak use I will make of them, explained in the following remark.

Remark 5.1 (Distinguishing vs. Witnessing)**.** Let us draw a distinction between a *distinguishing* use of the extra points and a *witnessing* use of them. Suppose that

[5]For a less extreme metaphysical hypothesis, consider the view that the laws of physics are metaphysically necessary. Even if true, we want to make logical distinctions beyond physical ones.

[6]If we were working with models for quantified epistemic logic, then they would be first-order structures that do not correspond to metaphysically possible worlds according to our intuitive understanding of the predicates, functions, and constants.

given the intended meaning of p and q, p is metaphysically equivalent to $p \wedge q$, but given the logical inequivalence of p and $p \wedge q$, we would like to allow the non-identity of $[\![p]\!]$ and $[\![p \wedge q]\!]$, so we include in our model(s) "impossible" points where p is true and q is false. This is a *distinguishing* use of these points. By contrast, one might wish to use impossible points to witness an agent's ignorance of some metaphysically necessary truth (e.g., Hersperus = Phosphorus), by including such points in alternative sets in $\mathbf{r}(P, w)$ and allowing them to be uneliminated by the agent. This is a *witnessing* use of the points. King [2007] and Chalmers [2011] discuss the witnessing use of points that do not correspond to metaphysically possible worlds, but here I will only use them for distinguishing. While theorists who use impossible points for witnessing may owe us a story about the nature of these entities, their relevance in inquiry, and how they are (un)eliminated, I take is that theorists who use impossible points only for distinguishing can adopt the stance of logical construction that I have taken.

Formally, the models to be introduced in §5.2 will be tuples containing (among other things) a pair $\mathbf{W} = \langle W, \{W_w\}_{w \in W} \rangle$, where W is a set of points and for each $w \in W$, we think of $W_w \subseteq W$ as the subset of points that are *possible* relative to w (so $w \in W_w$), with which we will interpret necessity formulas $\Box\varphi$. Those who do not wish to use impossible points for witnessing an agent's ignorance may assume

$$\begin{array}{rl} \textbf{r-possible} & \bigcup \mathbf{r}(P, w) \subseteq W_w \text{ and} \\ \textbf{u-possible} & \mathsf{u}(P, w) \subseteq W_w, \end{array}$$

where $\mathsf{u}(P, w)$ is the set of uneliminated alternatives for P, as in Chapter 3.

I will sometimes refer to W, with its associated propositional valuation V, as "logical space." However, one should not take this to mean that W and V must provide the same maximal space in all models. We include in a model as many points as necessary to make the logical and epistemic distinctions that we wish to capture in a given scenario, and we call this "logical space" for current modeling purposes; if we want to make more distinctions, we add more points to the model for a larger "logical space."[7] Eventually we may reach Kaplan's big model that contains any points needed

[7] Cf. Stalnaker's [1984, 58] point that there need not be a single domain of all possibilities.

to distinguish logically inequivalent formulas. For example, by adding points we may reach a model where for every set of atoms in At, there is a point in the model that satisfies exactly those atoms—in other words, a model containing all *state-descriptions* in the Carnap-inspired sense of Hendry and Pokriefka 1985—in which all inequivalent propositional formulas have distinct extensions.[8] More generally, by adding points we may reach a model in which for every set of formulas that is consistent according to epistemic logic **L**, there is a point in the model satisfying those formulas (as in a Henkin model used in the standard style of completeness proofs for epistemic logic). Formally, where $\mathsf{At} = \{p, q, r \dots\}$ is the set of sentence letters in our language, a model $\mathfrak{M}$ with domain W and valuation V may satisfy

$$\textbf{A-space} \quad \forall \mathsf{S} \subseteq \mathsf{At}\ \exists s \in W \colon \mathsf{S} = \{p \in \mathsf{At} \mid s \in V(p)\}.$$

More generally, given a logic **L** for our language $\mathcal{L}$, a model $\mathfrak{M}$ may satisfy

$$\textbf{L-space} \quad \forall\ \mathbf{L}\text{-consistent sets } \Sigma \text{ of formulas } \exists s \in W \colon \Sigma \subseteq \{\varphi \in \mathcal{L} \mid \mathfrak{M}, s \vDash \varphi\}.$$

For the purposes of modeling concrete epistemic scenarios, the standard practice in epistemic logic is to start with a "small model" that omits many logical possibilities, for two reasons. First, we might assume that many of these have already been eliminated as epistemic possibilities. Second, we can always add more points later to make finer distinctions between what the agent knows and does not know. The first reason marks a slight difference with the approach taken here, where we often like to show in our model points that have already been eliminated according to the u function. However, the second reason for adopting the "small model" approach still applies here.

The question of which valuations to include in our "logical space" W also depends on how we think of the letters $p, q, r \dots$ of our language.

Remark 5.2 (Statement Letters)**.** Burgess [2003, 154] distinguishes two ways of

[8] I say 'Carnap-inspired' because Carnap allows that certain state-descriptions may be excluded from consideration given the *meanings* of the atoms and analytic truths relating those meanings, whereas I am adopting what Ballarin [2005, 278] calls the "Wittgensteinian logical/combinatorial view" of state-descriptions, which includes all combinatorially consistent state description.

thinking about the letters in our propositional modal language:

> What is crucial is that one distinguish conceptually between statement letters thought of as representing *arbitrary* statements, and statement letters thought of as representing *independent atomic statements*. With the former, standard understanding, the restriction to a subset of all valuations ... need have nothing to do with a switch from logical to any kind of non-logical modalities, since it is required even for logical modalities, simply a reflection of the fact that with logically complex, logically interrelated statements instantiating the statement letters, not all combinations of truth values may be logically possible .... By contrast, [including all valuations] with the non-standard understanding of the role of statement letters [as representing independent atomic sentences] ... is appropriate for logical modalities; whereas [restricting to a subset of all valuations] with the same understanding ... will be appropriate for non-logical modalities, such as metaphysical modalities.

Burgess's distinction is important in connection with the **space** conditions above. If we think of statement letters as representing *independent atomic statements*,[9] then a

[9]Burgess [2003, 147-148] explicates this way of thinking as follows:

> The result of replacing the statement letters in a formula A with specific statements, such as "Snow is white" or "Snow is black" (with simultaneous replacement of logical symbols $\sim$, $\&$, $\vee$, and so on, by the logical operations of negation, conjunction, disjunction, and so on, that they are supposed to represent) I will call an *instantiation* of A .... Let us call the result of replacing the statement letters in a formula A by statements an *instantiation** if statement letters are replaced by *statements that are logically atomic* (or, to state explicitly once a qualification that will henceforth be tacitly understood, if not literally logically atomic, at least without further logical structure that can be represented only using whatever logical symbols one is using), and *with* [sic] *distinct sentence* [sic] *are instantiated by statements that are logically independent*. (Here n statements $\alpha_1, \ldots, \alpha_n$ are independent if all $2n$ combinations of truth values are possible.)
>
> A formula fully indicates the logical form of its instantiations* (insofar as it can be represented with the logical symbols one is using), but not of all its instantiations. For example, "Grass is green or snow is white" is an instantiation* of $p \vee q$, while "Grass is green or grass is not green" is an instantiation of $p \vee q$ that is not an instantiation* thereof. It is, rather, an instantiation* of $p \vee \sim p$ ....

condition like **A-space** makes sense. However, if we think of them as representing *arbitrary* statements, then we can have a full "logical space" without **A-space**, let alone **L-space**. Both ways of thinking are compatible with the framework of this chapter, and I will flag those places where the distinction matters.

5.1.4 Logical Structure

The example of multiple paths to knowing a disjunction in §5.1.1 assumes that the objects of knowledge have some internal *structure*. Can we implement this epistemological idea in our formal framework? Before answering this question, it helps to consider a related question. Those who accept the example of **contrast** failure for disjunctions in §5.1.2 often ask whether **contrast** should hold for claims that are atomic, e.g., *d is a duck*. Can we implement this idea in our formal framework? The answer to both questions is 'yes', and there multiple ways to do so. One way, discussed in §5.2.4, would be to take the first input of our new **r** function to be a *formula*, rather than a set. However, by taking advantage of the discussion in the previous section, we can implement the two ideas without changing the inputs to **r**.[10]

The two questions raise an interesting point of contrast between possible worlds semantics in the style of Lewis [1986] and Stalnaker [1984] and formal semantics in modal logic. In the pictures of Lewis and Stalnaker, the objects of knowledge are propositions—sets of worlds—and there is no sense in which one of these sets is "atomic" and another "complex." They are just sets. Hence there is no way to draw a distinction, using only such coarse-grained propositions, between knowing something atomic vs. knowing something complex. By contrast, in modal logic, in addition to sets of points we have a *valuation function* for sentence letters, which can be understood—following Burgess's distinction in Remark 5.2—as representing independent atomic sentences. This allow us to associate with sets of points some logical structure: some sets are definable by an atomic sentence or its negation, some sets are definable as the union of two distinct sets of the first kind, etc., and assuming

[10]Note added in ILLC version: in Holliday 2013b,c, I take the first input to **r** to be a structured proposition/formula. I then propose constraints on **r** such that logically equivalent propositions have essentially the same alternative sets, so mere syntactic differences do not matter. See §5.2.4 below.

A-space from §5.1.3, no set of the first "atomic" kind is also a set of the second "disjunctive" kind, etc. In this way, formal semantics in modal logic takes a step in the direction of structured propositions, away from the more coarse-grained pictures of Lewis and Stalnaker.[11] Hence we can draw the distinction between knowing something atomic vs. knowing something complex. To capture the idea that **contrast** holds for atomic propositions—and conjunctions thereof—one can require that the $\mathbf{r}$ function in a model (see Definition 5.1) satisfies the following for all propositions $P \subseteq W$ and worlds $w \in W$, where At is the set of atomic sentences in our language:

$$\textbf{A-contrast} \quad \forall p \in \mathsf{At} : \bigcup \mathbf{r}([\![\pm p]\!], w) \subseteq \overline{[\![\pm p]\!]};$$

$$\textbf{A-contrast}^{+} \quad \text{if } \exists p \in \mathsf{At} : P \subseteq [\![\pm p]\!], \text{ then } \bigcup \mathbf{r}(P, w) \subseteq \overline{P}.$$

There is another way to see that we can capture the idea that different paths to knowing depend on the logical structure of what is known, even when the inputs of our $\mathbf{r}$ function are sets. Suppose that to model a particular scenario, I construct a small model in which it happens to hold that $[\![p \vee q]\!] = [\![r]\!]$. Moreover, suppose that I claim there are multiple paths to knowing p *or* q, so there should be multiple alternative sets for $[\![p \vee q]\!]$. Hence $[\![r]\!]$ will have the same multiple alternative sets. If we are thinking of sentence letters like r as representing arbitrary sentences, then one may be fine with allowing multiple paths to knowing r and hence multiple alternative sets for $[\![r]\!]$. However, if we are understanding sentence letters like r as representing independent atomic sentences, then one may object to $[\![r]\!]$ having the same alternative sets as $[\![p \vee q]\!]$. But if so, then there is an simple solution: add a point to the model where r is true and $p \vee q$ false, or vice versa, so that $[\![p \vee q]\!] \neq [\![r]\!]$. (One can assume that this point is not in $\mathbf{r}(P, w)$ or $\mathsf{u}(P, w)$ for any P and w.) Having pulled apart the extensions in this way, one can then assign to them different alternative sets.

Of course, if the first input to the $\mathbf{r}$ function is a set, then we cannot assign different alternatives sets to logical equivalents, since these always have the same extensions. Hence this approach assumes that what matters for the multiple paths to knowledge

[11] For the formal counterpart of the Lewis and Stalnaker pictures, see Halpern 1999 on purely set-theoretic approaches to epistemic logic (i.e., with no syntax) as used in economics.

is *logical* structure, rather than finer-grained *syntactic* structure. To represent views according to which logical equivalents can have different alternative sets, we must change the inputs of **r**. I will discuss the relation between these approaches in §5.2.4.

5.1.5 Logical Closure

Given our distinction between logical space W and the metaphysical space W_w for a given world w, we can represent in our formalism a fundamental distinction between two types of closure for ideally astute logicians (IALs): closure under *logical consequence* (or *equivalence*) and closure under (known) *strict implication* (or *bi-implication*). In the epistemology literature, authors often play fast and loose with the notion of logical consequence and equivalence. For example, the idea that that *the animals in the pen aren't cleverly disguised mules* is a "clear logical consequence" of *the animals in the pen are zebras* (Vogel 1990, 40) confuses logic and zoology. Similarly, the idea that one could know of a zebra that "its being a zebra is *a priori* (or logically) equivalent to its being a zebra and not a painted mule" [Adams et al., 2012] involves the same zoo-logical confusion.[12] To state the obvious:

> 'z is a zebra **or** z is not a painted mule' is a logical consequence of 'z is a zebra';
>
> 'z is not a painted mule' is a logical consequence of 'z is not a painted mule **and** z is a zebra'.

By contrast:

> 'z is a zebra **and** z is not a painted mule' is *not* a logical consequence of 'z is a zebra';
>
> 'z is not a painted mule' is *not* a logical consequence of 'z is a zebra'.

At best, the latter two are cases of strict implication.

Having distinguished logic from zoology, let us state *closure under logical consequence* as the principle that if ψ is a logical consequence of φ, so $\varphi \to \psi$ is a *logical truth*, then the IAL knows φ only if she knows ψ. I will also call this

[12]Moreover, as Hawthorne [2004a, 41n99] notes, it is at best a necessary a posteriori truth that zebras are not mules.

- single-premise *logical* closure, represented by the rule

$$\text{RM } \frac{\varphi \to \psi}{K\varphi \to K\psi}.$$

Since I assume that IALs know all logic, I will not bother to distinguish between closure under logical consequence and closure under *known* logical consequence. By contrast, since I will not always assume that IALs know all metaphysics, I will distinguish between closure under strict implication and closure under *known* strict implication. Formally, we distinguish single-premise logical closure from:

- closure under known *strict implication*,

$$(K\varphi \wedge K\Box(\varphi \to \psi)) \to K\psi;$$

- closure under *strict implication*,

$$(K\varphi \wedge \Box(\varphi \to \psi)) \to K\psi;$$

- closure under known *material implication*,

$$\text{K } (K\varphi \wedge K(\varphi \to \psi)) \to K\psi.$$

Finally, let us distinguish single-premise logical closure from *multi-premise* logical closure, the principle that if ψ is a logical consequence of $\{\varphi_1, \ldots, \varphi_n\}$, so $\varphi_1 \wedge \cdots \wedge \varphi_n \to \psi$ is a logical truth, then the IAL knows $\varphi_1, \ldots, \varphi_n$ only if she knows ψ:

- multi-premise logical closure,

$$\text{RK } \frac{\varphi_1 \wedge \cdots \wedge \varphi_n \to \psi}{K\varphi_1 \wedge \cdots \wedge K\varphi_n \to K\psi}.$$

If we assume that all logical truths are metaphysically necessary and known by IALs,[13] and that whatever is metaphysically necessary is true,[14] then the five closure principles above are listed in order of *increasing deductive strength*, except for K and RK, which have the same deductive power (assuming IALs know all logical truths). The last point means that anyone who rejects closure under material implication must also reject multi-premise logical closure, an important point to which we will return.

Since our models in this chapter will contain both a logical space W and a metaphysical space W_w for a given $w \in W$, we will be able to semantically distinguish single premise *logical* closure not only from closure under known implication, as we could in Chapter 3, but also from the metaphysical closure principles.

Analogous distinctions apply to equivalence, strict bi-implication, and bi-implication:

- closure under *logical equivalence*,

$$\text{RE } \frac{\varphi \leftrightarrow \psi}{K\varphi \leftrightarrow K\psi}.$$

- closure under known *strict bi-implication*,

$$(K\varphi \wedge K\Box(\varphi \leftrightarrow \psi)) \rightarrow K\psi;$$

- closure under *strict bi-implication*,

$$(K\varphi \wedge \Box(\varphi \leftrightarrow \psi)) \rightarrow K\psi;$$

- closure under known *material bi-implication*,

$$(K\varphi \wedge K(\varphi \leftrightarrow \psi)) \rightarrow K\psi.$$

[13] That is, the necessitation rules for $\Box$ and K.
[14] That is, the T axiom $\Box\alpha \rightarrow \alpha$.

5.1.6 Main Claims

With the distinctions of §5.1.5, I can now state the main claims of this chapter. First, the Multipath Picture fixes the flaws of Fallibilism 1.0 discussed in Chapter 4:

Claim 5.1 (The Three Problems Solved). In Fallibilism 1.0, we were forced to either admit vacuous knowledge or give up even special cases of single-premise logical closure such as $K(\varphi \wedge \psi) \rightarrow K\varphi$ and $K\varphi \rightarrow K(\varphi \vee \psi)$. Yet in the Multipath Picture for Fallibilism 2.0, we can reject vacuous knowledge and retain all single-premise logical closure principles, thereby solving the twin problems of Vacuous Knowledge and Containment without resorting to Knowledge Inflation. In §5.4, I will give a new account of deduction as involving knowledge extension rather than inflation.

If Claim 5.1 is correct, then it follows that one of the most serious concerns about fallibilism without full closure—that it will force us into the extreme closure failures that plagued the subjunctivist-flavored theories in Chapter 2—has been eliminated. In addition to establishing this claim in defense of fallibilism without full closure, I will argue for the following claims against fallibilism with full closure in §5.3:

Claim 5.2 (Against Implication Closure). Fallibilist should reject the idea that closure under known implication is valid. That principle (or equivalently, multi-premise logical closure) still saddles fallibilists with the Problem of Vacuous Knowledge.

Claim 5.3 (Against Strict Implication Closure). Fallibilists should reject the idea that closure under (known) strict implication/bi-implication is valid. That principle either forces fallibilists into the Problem of Vacuous Knowledge or forces fallibilists to give up even $K(\varphi \wedge \psi) \rightarrow K\varphi$, reinstating the Problem of Containment. In §6.1.2, I will explain what I take to be the mistake in accepting strict bi-implication closure.

Fallibilists may accept the Multipath Picture of Knowledge without accepting Claims 5.2 or 5.3. Roughly speaking, fallibilists who insist on full closure—and are therefore committed to vacuous knowledge—will be committed to *less* vacuous knowledge in the Multipath Picture than they were in Fallibilism 1.0. However, it is still too much vacuous knowledge for me, so I will argue firmly for Claims 5.2 and 5.3.

5.2 Multipath Alternatives Models

Let us now develop the Multipath Picture formally, starting with our new models.

Definition 5.1 (MA Model)**.** A *multipath alternatives* model is a tuple $\mathfrak{M}$ of the form $\langle \mathbf{W}, \mathsf{u}, \mathbf{r}, V\rangle$ where $\mathbf{W} = \langle W, \{W_w\}_{w\in W}\rangle$ with W a non-empty set and $w \in W_w \subseteq W$; $\mathsf{u}\colon \mathcal{P}(W) \times W \to \mathcal{P}(W)$, $\mathbf{r}\colon \mathcal{P}(W) \times W \to \mathcal{P}(\mathcal{P}(W))$, and $V\colon \mathsf{At} \to \mathcal{P}(W)$.

As in §5.1.3, W is logical space and W_w is the set of worlds that are (metaphysically) possible relative to w. (One may assume standard constraints on this notion of possibility, e.g., requiring that if $v \in W_w$, then $W_v = W_w$, but none of this will matter for our purposes.) As in Chapter 3, $\mathsf{u}(P, w)$ is the set of alternatives (possibilities) that are *uneliminated as alternatives for* P by the agent in w. However, in contrast to $\mathsf{r}(P, w)$ from Chapter 3, $\mathbf{r}(P, w)$ is not a single set of alternatives that the agent must rule out in order to know P in w. Rather, $\mathbf{r}(P, w)$ is a set *of sets* of alternatives such that in order to know P in w, the agent must rule out all of the alternatives in at least *one of* these sets. This is precisely the content of the following truth definition.

Definition 5.2 (Truth in a MA Model)**.** Given a model $\mathfrak{M} = \langle \mathbf{W}, \mathsf{u}, \mathbf{r}, V\rangle$ with $w \in W$ and a formula φ in the epistemic-alethic language (recall §2.9.2), we define $\mathfrak{M}, w \vDash \varphi$ as follows (with propositional cases as usual):

$$\begin{array}{lll} \mathfrak{M}, w \vDash \Box\varphi & \text{iff} & W_w \subseteq [\![\varphi]\!]^{\mathfrak{M}}; \\ \mathfrak{M}, w \vDash K\varphi & \text{iff} & \exists A \in \mathbf{r}([\![\varphi]\!]^{\mathfrak{M}}, w)\colon A \cap \mathsf{u}([\![\varphi]\!]^{\mathfrak{M}}, w) = \emptyset, \end{array}$$

where $[\![\varphi]\!]^{\mathfrak{M}} = \{v \in W \mid \mathcal{M}, v \vDash \varphi\}$.

Observe that we can assume without loss of generality that all models are *non-redundant* in the sense that for all $P \subseteq W$, $w \in W$, and $A, B \in \mathbf{r}(P, w)$: $A \not\subseteq B$.

As in Chapter 3, with no constraints on $\mathbf{r}$ or u we have the following result.

Proposition 5.1 (Completeness of E)**.** **E** is sound and complete for the class of all MA models.

5.2.1 The Five Postulates

Rather than studying possible constraints on **r** one-by-one, as in Chapter 3, I will go straight to my own theory of the **r** function, consisting of the Five Postulates in Definition 5.3. Included among these postulates are the analogues in the Multipath Picture of noVK and cover from Chapters 3 and 4, now called **noVK** and **cover**. In §5.2.3, I will show the consistency of the Five Postulates together with **fallibilism**, defined below. I have stated strong postulates and more postulates than are needed to resolve the problems of Vacuous Knowledge and Containment, since this makes the consistency result a stronger result. Otherwise one may worry that as soon as we add to **noVK** and **cover** other constraints, we will find an inconsistency. Moreover, the reader may rest assured that any weakening of the postulates is also consistent.

Definition 5.3 (Five Postulates)**.** An MA model $\mathfrak{M} = \langle \mathbf{W}, \mathsf{u}, \mathbf{r}, V \rangle$ satisfies the Five Postulates if and only if for all $P \subseteq W$ and $w \in W$:

1. (**r-RofA**) if $w \in \overline{P}$, then $\mathbf{r}(P, w) = \emptyset$;
   Read: if w is a not-P world, then there is no path to knowing P in w.

2. (**enough**) if $w \in P$, then $\exists A \in \mathbf{r}(P, w)$: $A \subseteq \overline{P}$;
   Read: in order to know P in w, it is sufficient that one eliminates *all* not-P possibilities, which the existence of such an alternative set A for P guarantees. See Fig. 5.4.[15]

3. (**noVK**) if $P \neq W_w$, then $\emptyset \notin \mathbf{r}(P, w)$;
   Read: if P is contingent, then coming to know P in w requires eliminating some possibilities, in which case the empty set cannot be an alternative set for P.[16]

4. (**overlap**) $\forall A \in \mathbf{r}(P, w)$: if $A \cap P \neq \emptyset$, then $\exists Q \subsetneq P$: $A \in \mathbf{r}(Q, w)$;
   Read: if an alternative set A for P *overlaps* with P—so there is a path to knowing P that involves eliminating P-possibilities—then this is because there

[15] Together **enough** and **r-possible** guarantee the validity of $\Box\varphi \rightarrow K\varphi$, so agents know all metaphysical necessities, as in standard possible worlds models without impossible points. This raises interesting issues, but my focus here is on knowledge of contingent empirical propositions.

[16] Those who reject **r-possible** may wish to restate **noVK** as: if $P \neq W$, then $\emptyset \notin \mathbf{r}(P, w)$.

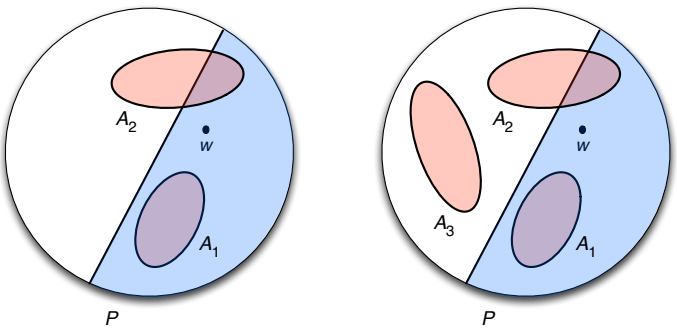

Figure 5.4: enough violated (left) vs. satisfied (right)

is some Q that is logically stronger than P such that eliminating all of A is a path to knowing Q. See Fig. 5.5.[17]

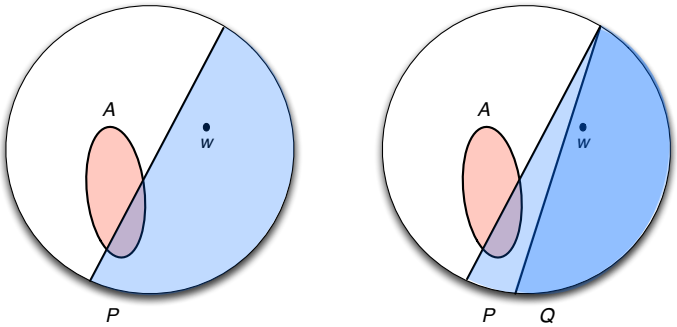

Figure 5.5: overlap visualized

5. (**cover**) if $Q \subseteq P$, then $\forall B \in \mathbf{r}(Q, w)\ \exists A \in \mathbf{r}(P, w)$: $A \subseteq B$.
   Read: if Q excludes as much of logical space as P does, then any path to knowing Q by eliminating possibilities covers a path to knowing P.

At this point, the Five Postulates should be largely self-explanatory. **r-RofA** is an

[17]Strictly speaking, Fig. 5.5 reflects a stronger statement of **overlap**, which is also consistent with the other postulates:

(**overlap**$^+$) $\forall A \in \mathbf{r}(P, w)$: if $A \cap P \neq \emptyset$, then $\exists Q \subsetneq P$: $A \cap Q = \emptyset$ and $A \in \mathbf{r}(Q, w)$.

Read: if an alternative set A for P *overlaps* with P—so there is a path to knowing P that involves eliminating P-possibilities—then this is because there is some Q that is logically stronger than P (and does not overlap with A) such that eliminating all of A is a path to knowing Q.

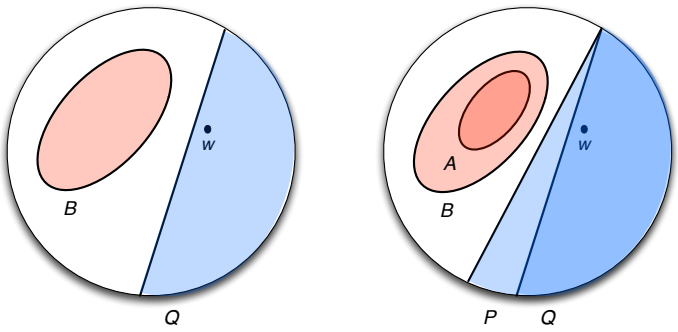

Figure 5.6: cover visualized

analogue of r-RofA from Chapter 3.[18] I have already argued for **enough** (though not under this name) in §4.2.1 and for **noVK** in §4.1. The one new postulate is **overlap**, which I alluded to in §5.1.2. This postulate generalizes the example from §5.1.2 into a rule about when **contrast** fails: think of P in the statement of **overlap** as the extension of $\alpha \vee \beta$ and the stronger proposition Q as the extension α. The point of adding this postulate, besides its plausibility, is to show that even if we strongly constrain failures of **contrast** as in **overlap**, we can still consistently satisfy all of the other postulates. One may worry that with no constraints on failures of **contrast**, anything goes, so no wonder we have consistency. In addition to **overlap**, we can consistently add the constraint on failures of **contrast** mentioned in §5.1.4: to capture the idea that **contrast** holds for knowing atomic propositions (and conjunctions thereof), one can assume **A-contrast** (and **A-contrast**$^+$) from §5.1.4

Let us now consider the **cover** postulate. What is crucial to observe is that the antecedent of **cover**, $Q \subseteq P$, is a statement of set inclusion in logical space W, not in metaphysical space W_w. Hence the following principle is *not* a consequence of **cover**: if Q strictly implies P, so $Q \cap W_w \subseteq P$, then any path to knowing Q by eliminating possibilities covers a path to knowing P, which implies (unlike **cover**) that if Q and P are metaphysically equivalent, so $Q \cap W_w = P \cap W_w$, then knowing Q requires eliminating the same possibilities as knowing P, i.e., $\mathbf{r}(Q, w) = \mathbf{r}(P, w)$.[19] In my view, this

[18] An alternative definition of **r-RofA**, which is even more analogous to r-RofA, requires that if $w \in \overline{P}$, then $w \in \bigcap \mathbf{r}(P, w)$, which would serve our purposes just as well given u-RofA.

[19] In other words, the principle

(**M-cover**) if $Q \cap W_w \subseteq P$, then $\forall B \in \mathbf{r}(Q, w)\ \exists A \in \mathbf{r}(P, w)$: $A \subseteq B$,

principle—as well as the weaker version for *known* strict implication/bi-implication—should be anathema to fallibilists, for reasons explained in §5.3. However, for now let us observe the following correspondence results, recalling §3.2.3.[20]

Proposition 5.2 (Correspondence)**.** Let $\langle \mathbf{W}, \mathbf{r} \rangle$ be an **r**-frame.

1. $K(\varphi \wedge \psi) \rightarrow (K\varphi \wedge K\psi)$ is valid on $\langle \mathbf{W}, \mathbf{r} \rangle$ relative to models satisfying $\mathsf{RO}_{\exists\forall}$ iff **r** satisfies **cover**.

2. $(K\varphi \wedge \Box(\varphi \rightarrow \psi)) \rightarrow K\psi$ is valid on $\langle \mathbf{W}, \mathbf{r} \rangle$ relative to models satisfying $\mathsf{RO}_{\exists\forall}$ iff **r** satisfies

$$(\textbf{M-cover}) \;\; \text{if } Q \cap W_w \subseteq P, \text{ then } \forall B \in \mathbf{r}(Q, w)\; \exists A \in \mathbf{r}(P, w)\colon A \subseteq B.$$

3. $(K\varphi \wedge \Box(\varphi \leftrightarrow \psi)) \rightarrow K\psi$ is valid on $\langle \mathbf{W}, \mathbf{r} \rangle$[21] relative to models satisfying $\mathsf{RO}_{\exists\forall}$ iff **r** satisfies

$$(\textbf{M-equiv}) \;\; \text{if } Q \cap W_w = P \cap W_w, \text{ then } \mathbf{r}(Q, w) = \mathbf{r}(P, w).$$

4. $(K\varphi \wedge K\psi) \rightarrow K(\varphi \wedge \psi)$ is valid on $\langle \mathbf{W}, \mathbf{r} \rangle$ relative to models satisfying $\mathsf{RO}_{\exists\forall}$ iff **r** satisfies the following **combine** condition:

$$\forall P, P' \subseteq W\; \forall Q \in \mathbf{r}(P, w)\; \forall Q' \in \mathbf{r}(P', w)\; \exists S \in \mathbf{r}(P \cap P', w)\colon S \subseteq Q \cup Q'.$$

In §5.3 and §6.1.2 I will explain why fallibilists should reject the **M-cover** ('**M**' for metaphysical), **M-equiv**, and **combine** assumptions, as in Claims 5.3 and 5.2.

---

together with $Q \cap W_w = P \cap W_w$, implies $\mathbf{r}(Q, w) = \mathbf{r}(P, w)$, assuming as after Definition 5.2 that these sets are non-redundant in the sense that we never have $A, B \in \mathbf{r}(Y, w)$ such that $A \subsetneq B$. Suppose for reductio that there is $X \in \mathbf{r}(Q, w)$ but $X \notin \mathbf{r}(P, w)$. Then by **M-cover** and the fact that $Q \cap W_w \subseteq P$, there is some $X' \in \mathbf{r}(P, w)$ such that $X' \subsetneq X$. Then by **M-cover** and the fact that $P \cap W_w \subseteq Q$, there is some $X'' \in \mathbf{r}(Q, w)$ such that $X'' \subseteq X'$. But then $X'' \subsetneq X$, contradicting the assumption that we do not have $A, B \in \mathbf{r}(Q, w)$ such that $A \subsetneq B$.

[20] The $\mathsf{RO}_{\exists\forall}$ condition is defined for MA models exactly as it was for SA models in §3.2.1.

[21] Here we assume that **r** is non-redundant as after Definition 5.2 and in footnote 19.

In line with the above correspondence results, we can prove completeness theorems as in Chapter 3. For our purposes, the most important is the following. Recall the nomenclature for modal axioms: M is $K(\varphi \wedge \psi) \to (K\varphi \wedge K\psi)$ and T is $K\varphi \to \varphi$.

Proposition 5.3 (Completeness of EMNT)**.** **EMNT** is sound and complete for the class of MA models satisfying the Five Postulates, $\mathsf{RO}_{\exists\forall}$, and u-RofA.

Hence if we add to the Five Postulates on the $\mathbf{r}$ function a very simple theory of the u function with $\mathsf{RO}_{\exists\forall}$ and u-RofA, then we obtain EMNT as our static epistemic logic for reasoning about what an ideally astute logician knows (for a fixed context). It is easy to check that EMNT is equivalent to propositional logic plus the rule

$$\text{RM } \frac{\varphi \to \psi}{K\varphi \to K\psi}$$

and the axioms $K\top$ and $K\varphi \to \varphi$. So with the simply theory of u, knowledge implies truth, the ideally astute logician knows all logical truths, and the ideally astute logician knows all logical consequences of each proposition she knows. However, in §5.4 I will replace the simple theory of u with a dynamic theory that models the epistemic effect of "putting two and two together," necessary for those of us who are not yet ideally astute logicians. We will then replace RM with a dynamic rule.

All along my arguments have been motivated by fallibilism, so it is time to state what fallibilism minimally amounts to in the Multipath Picture, using Definition 5.4.

Definition 5.4 (Fallibilism)**.** A MA model $\mathfrak{M} = \langle \mathbf{W}, \mathsf{u}, \mathbf{r}, V \rangle$ satisfies **fallibilism** iff

$$\exists P \subseteq W \, \exists w \in W \, \exists A \in \mathbf{r}(P, w) \colon w \in P \text{ and } A \subsetneq \overline{P}.$$

In other words, fallibilism implies that there is some world w where an agent can come to know a proposition P by ruling out a *strict subset* of the not-P possibilities. Stronger versions of fallibilism—according to which there are many such w and P—are also compatible with the Five Postulates, as should be clear from §5.2.3.

5.2.2 Lewis and Nozick in MA Models

Having introduced all of the conditions on the $\mathbf{r}$ function to be considered in this chapter, it helps to see how the pictures of Lewis and Nozick look in our MA models—and how they violate the Five Postulates. Given an RA model $\mathcal{M} = \langle W, \rightarrow\!\!\!\triangleright, \preceq, V\rangle$ from §2.4, as in Fig. 5.7, we can define an MA model $\mathfrak{M} = \langle \mathbf{W}, \mathsf{u}, \mathbf{r}, V\rangle$ for Lewis by taking each W_w in the MA model to be the field of $\preceq_w$ from the RA model and by defining $\mathbf{r}(P, w) = \{\{\mathrm{Min}_{\preceq_w}(W) \cap \overline{P}\}\}$. Of course, Lewis's picture does not take advantage of the multiple paths to knowledge in the Multipath Picture, so $\mathbf{r}(P, w)$ only contains a single alternative set. For u, we define $\mathsf{u}(P, w) = \{v \in \overline{P} \mid w \rightarrow\!\!\!\triangleright\, v\}$. In this case, while MA models contain a function u instead of a relation $\rightarrow\!\!\!\triangleright$, we can draw the arrows in our diagrams and recover the function by the definition just given.

In Fig. 5.7, which is the same as Fig. 2.1 from §2.4 except for the names of atoms, think of p as some mundane proposition and s as an incompatible skeptical hypothesis (in particular, a world-side skeptical hypothesis in the sense of §6.2.3). In the actual world w_1, p is true, but there is also a "relevant" or "close" world w_2 in which p is false, which is strictly more relevant or closer than the skeptical world w_3 in which p is false and the skeptical hypothesis s is true. Depending on the specific choices for p and s, the world w_4 in which p and s are both true may be metaphysically impossible or only impossible in a weaker sense, e.g., physically impossible.

Remark 5.3 (Strong vs. Strict Skeptical Counter-Hypotheses)**.** In the rest of this section and in §5.2.3, I treat the case where s is what I call a (merely) *strong* skeptical counter-hypothesis to p, in the sense that $p \rightarrow \neg s$ is necessary in a weaker sense than metaphysical necessity, as was implicitly assumed in the medical diagnosis Example 1.1 in Chapter 2 (where p was c and s was x).[22] Hence we let $w_4 \in W_{w_1}$. In §5.3, I treat the case where s is what I call a *strict* counter-hypothesis to p, in the sense that $p \rightarrow \neg s$ is necessary in the strongest sense. The birdwatching Example 1.2 is such a case, assuming it is metaphysically impossible for something to be a Gadwall and a Siberian Grebe (cf. Stroud [1984, 25]: "a goldfinch simply could not be a canary").

[22] As explained in Example 1.1, given actual human biology, x confers immunity to c, so $c \rightarrow \neg x$ is "biologically necessary." But we can suppose that this is a contingent biological truth; if human biology were slightly different in various ways, then x would not confer immunity to c.

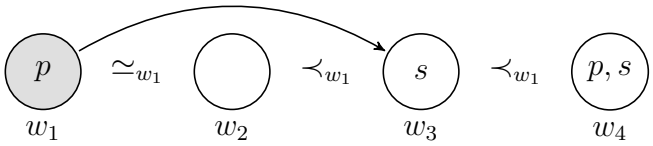

Figure 5.7: RA model from §2.4 (partially drawn, reflexive loops omitted)

Example 5.1 (Lewisian MA Model)**.** Deriving the **r** function for our Lewisian MA model from the RA model in Fig. 5.7, we have the result in Table 5.1, where I have grayed out the information for propositions false at w_1. Fig. 5.8 displays the same result in a color-coded graphical form. Each row in Fig. 5.8 displays a proposition, *the set of worlds outlined in blue*, along with the single Lewisian alternative set for that proposition consisting of the *red worlds*. (Each non-empty alternative set in this model happens to contain a single world, but of course that is not required.) Note the conspicuous violations of the **noVK** postulate in rows four, six, and seven.

Let us now construct a Nozickian MA model based on the RA model in Fig. 5.7. For Nozick, strictly we should start with the CB model in Fig. 2.2 of §2.5 and take into account the role of belief in N-semantics when defining the u function as in §3.3.2, but for simplicity I will start with the Nozick-like picture of D-semantics over RA models from §2.4. The result for the **r** function will be exactly the same. For our MA model we define $\mathbf{r}(P, w) = \{\{\text{Min}_{\preceq_w}(\overline{P})\}\}$. As in Example 5.1, since we are starting with a theory in the framework of Fallibilism 1.0, we are not yet taking advantage of the multiple paths to knowledge available in the Multipath Picture.

Example 5.2 (Nozickian MA Model)**.** Deriving the **r** function for our Nozickian MA model from the RA model in Fig. 5.7, we have the result in Table 5.2, where I have grayed out the information for propositions that are false at w_1. Fig. 5.9 displays the same result in a color-coded graphical form. Each row in Fig. 5.8 displays a proposition, *the set of worlds outlined in blue*, along with the single Nozickian alternative set for that proposition consisting of the *red worlds*. (Each non-empty alternative set in this model happens to contain a single world, but of course that is not required.) Note the conspicuous violations of the **cover** postulate in rows four

$$\mathsf{U}(w_1) = \{w_1, w_3\}$$
$$\mathsf{u}(P, w_1) = \mathsf{U}(w_1) \cap \overline{P}$$

$\mathbf{r}([\![p \wedge \neg p]\!], w_1) =$	$\mathbf{r}(\emptyset, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p \wedge s]\!], w_1) =$	$\mathbf{r}(\{w_4\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p \wedge \neg s]\!], w_1) =$	$\mathbf{r}(\{w_1\}, w_1)$	$= \{\{w_2\}\}$	$\mathfrak{M}, w_1 \vDash K(p \wedge \neg s)$
$\mathbf{r}([\![\neg p \wedge s]\!], w_1) =$	$\mathbf{r}(\{w_3\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![\neg p \wedge \neg s]\!], w_1) =$	$\mathbf{r}(\{w_2\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p]\!], w_1) =$	$\mathbf{r}(\{w_1, w_4\}, w_1)$	$= \{\{w_2\}\}$	$\mathfrak{M}, w_1 \vDash Kp$
$\mathbf{r}([\![s]\!], w_1) =$	$\mathbf{r}(\{w_3, w_4\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p \leftrightarrow s]\!], w_1) =$	$\mathbf{r}(\{w_2, w_4\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p \leftrightarrow \neg s]\!], w_1) =$	$\mathbf{r}(\{w_1, w_3\}, w_1)$	$= \{\{w_2\}\}$	$\mathfrak{M}, w_1 \vDash K(p \leftrightarrow \neg s)$
$\mathbf{r}([\![\neg s]\!], w_1) =$	$\mathbf{r}(\{w_1, w_2\}, w_1)$	$= \{\emptyset\}$	$\mathfrak{M}, w_1 \vDash K\neg s$
$\mathbf{r}([\![\neg p]\!], w_1) =$	$\mathbf{r}(\{w_2, w_3\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p \vee s]\!], w_1) =$	$\mathbf{r}(\{w_1, w_3, w_4\}, w_1)$	$= \{\{w_2\}\}$	$\mathfrak{M}, w_1 \vDash K(p \vee s)$
$\mathbf{r}([\![p \vee \neg s]\!], w_1) =$	$\mathbf{r}(\{w_1, w_2, w_4\}, w_1)$	$= \{\emptyset\}$	$\mathfrak{M}, w_1 \vDash K(p \vee \neg s)$
$\mathbf{r}([\![p \rightarrow s]\!], w_1) =$	$\mathbf{r}(\{w_2, w_3, w_4\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p \rightarrow \neg s]\!], w_1) =$	$\mathbf{r}(\{w_1, w_2, w_3\}, w_1)$	$= \{\emptyset\}$	$\mathfrak{M}, w_1 \vDash K(p \rightarrow \neg s)$
$\mathbf{r}([\![p \vee \neg p]\!], w_1) =$	$\mathbf{r}(W, w_1)$	$= \{\emptyset\}$	$\mathfrak{M}, w_1 \vDash K(p \vee \neg p)$

Table 5.1: partial representation of the Lewisian MA model from Example 5.1

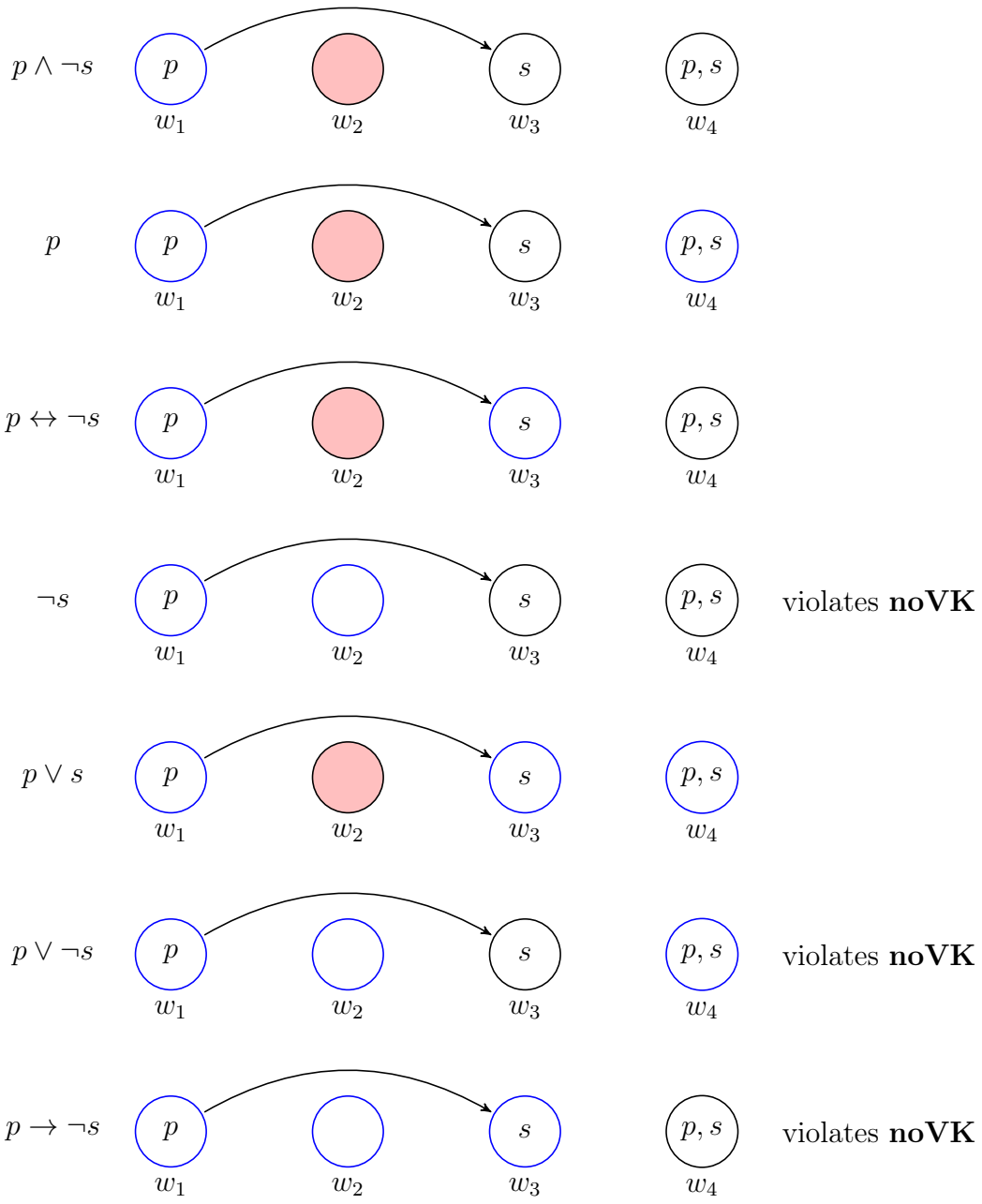

Figure 5.8: partial representation of the Lewisian MA model from Example 5.1

$$\mathsf{U}(w_1) = \{w_1, w_3\}$$
$$\mathsf{u}(P, w_1) = \mathsf{U}(w_1) \cap \overline{P}$$

$\mathbf{r}([\![p \wedge \neg p]\!], w_1) =$	$\mathbf{r}(\emptyset, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p \wedge s]\!], w_1) =$	$\mathbf{r}(\{w_4\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p \wedge \neg s]\!], w_1) =$	$\mathbf{r}(\{w_1\}, w_1)$	$= \{\{w_2\}\}$	$\mathfrak{M}, w_1 \vDash K(p \wedge \neg s)$
$\mathbf{r}([\![\neg p \wedge s]\!], w_1) =$	$\mathbf{r}(\{w_3\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![\neg p \wedge \neg s]\!], w_1) =$	$\mathbf{r}(\{w_2\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p]\!], w_1) =$	$\mathbf{r}(\{w_1, w_4\}, w_1)$	$= \{\{w_2\}\}$	$\mathfrak{M}, w_1 \vDash Kp$
$\mathbf{r}([\![s]\!], w_1) =$	$\mathbf{r}(\{w_3, w_4\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p \leftrightarrow s]\!], w_1) =$	$\mathbf{r}(\{w_2, w_4\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p \leftrightarrow \neg s]\!], w_1) =$	$\mathbf{r}(\{w_1, w_3\}, w_1)$	$= \{\{w_2\}\}$	$\mathfrak{M}, w_1 \vDash K(p \leftrightarrow \neg s)$
$\mathbf{r}([\![\neg s]\!], w_1) =$	$\mathbf{r}(\{w_1, w_2\}, w_1)$	$= \{\{w_3\}\}$	$\mathfrak{M}, w_1 \nvDash K\neg s$
$\mathbf{r}([\![\neg p]\!], w_1) =$	$\mathbf{r}(\{w_2, w_3\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p \vee s]\!], w_1) =$	$\mathbf{r}(\{w_1, w_3, w_4\}, w_1)$	$= \{\{w_2\}\}$	$\mathfrak{M}, w_1 \vDash K(p \vee s)$
$\mathbf{r}([\![p \vee \neg s]\!], w_1) =$	$\mathbf{r}(\{w_1, w_2, w_4\}, w_1)$	$= \{\{w_3\}\}$	$\mathfrak{M}, w_1 \nvDash K(p \vee \neg s)$
$\mathbf{r}([\![p \rightarrow s]\!], w_1) =$	$\mathbf{r}(\{w_2, w_3, w_4\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p \rightarrow \neg s]\!], w_1) =$	$\mathbf{r}(\{w_1, w_2, w_3\}, w_1)$	$= \{\{w_4\}\}$	$\mathfrak{M}, w_1 \vDash K(p \rightarrow \neg s)$
$\mathbf{r}([\![p \vee \neg p]\!], w_1) =$	$\mathbf{r}(W, w_1)$	$= \{\emptyset\}$	$\mathfrak{M}, w_1 \vDash K(p \vee \neg p)$

Table 5.2: partial representation of the Nozickian MA model from Example 5.2

(relative to row one), six (relative to rows one and four), and seven (relative to rows one and four). The following is the absurd consequence of the first violation of **cover**:

- While $p \wedge \neg s$ is a strictly stronger proposition than $p \vee \neg s$, in the Nozickian picture of Fig. 5.9, coming to know $p \vee \neg s$ requires additional epistemic work compared to coming to know $p \wedge \neg s$, so the agent knows $p \wedge \neg s$ but not $p \vee \neg s$.

Having seen how the standard "single path" pictures violate the Five Postulates in our MA models, in the next section we will see how taking advantage of multiple paths to knowledge establishes the consistency of the Five Postulates with fallibilism.

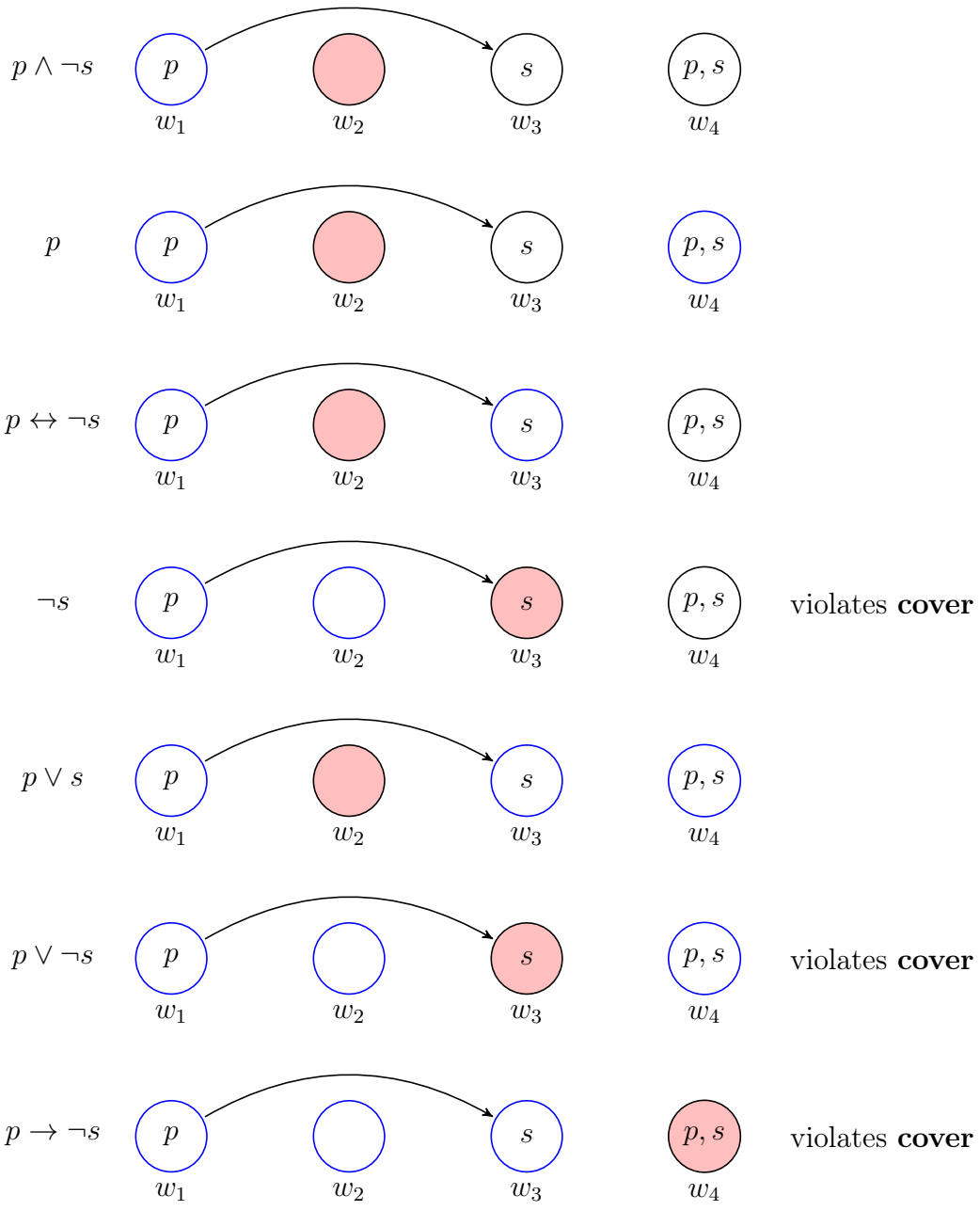

Figure 5.9: partial representation of the Nozickian MA model from Example 5.2

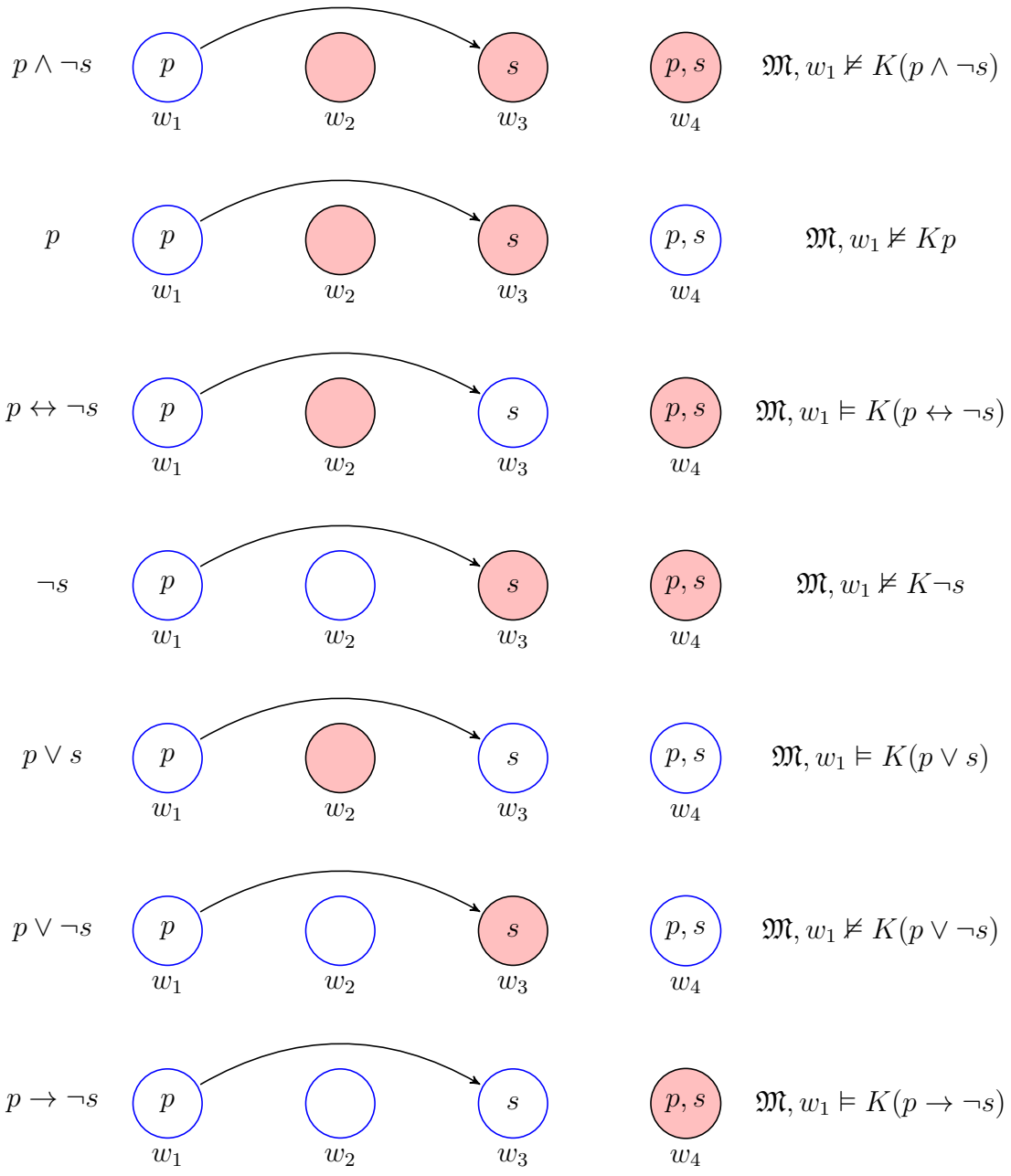

Figure 5.10: infallibilist MA model based on the RA model in Fig. 5.7

5.2.3 Consistency of the Postulates

It is now time to solve two of the problems of Fallibilism 1.0—the Problem of Vacuous Knowledge and the Problem of Containment—in one fell swoop. What prevented such a solution in the framework of Fallibilism 1.0 was the following impossibility result.

Proposition 4.1 (Impossibility I)**.** There is no SA model satisfying the following:

contrast	$\mathsf{r}(P,w) \subseteq \overline{P}$;
fallibilism	$\exists P \subseteq W\colon \overline{P} \not\subseteq \mathsf{r}(P,w)$;
noVK	if $P \neq W$, then $\mathsf{r}(P,w) \neq \emptyset$;
cover	if $P \subseteq Q$, then $\mathsf{r}(Q,w) \subseteq \mathsf{r}(P,w)$.

In contrast to Proposition 4.1, Proposition 5.4 finally delivers some good news.

Proposition 5.4 (Consistency)**.** The following are consistent:

1. (**r-RofA**) if $w \in \overline{P}$, then $\mathbf{r}(P,w) = \emptyset$;
2. (**enough**) if $w \in P$, then $\exists A \in \mathbf{r}(P,w)\colon A \subseteq \overline{P}$;
3. (**noVK**) if $P \neq W_w$, then $\emptyset \notin \mathbf{r}(P,w)$;
4. (**overlap**) $\forall A \in \mathbf{r}(P,w)$: if $A \cap P \neq \emptyset$, then $\exists Q \subsetneq P\colon A \in \mathbf{r}(Q,w)$;
5. (**cover**) if $Q \subseteq P$, then $\forall B \in \mathbf{r}(Q,w)\ \exists A \in \mathbf{r}(P,w)\colon A \subseteq B$;
6. (**r-possible**) $\bigcup \mathbf{r}(P,w) \subseteq W_w$;
7. (**A-space**) $\forall \mathsf{S} \subseteq \mathsf{At}\ \exists s \in W\colon \mathsf{S} = \{p \in \mathsf{At} \mid s \in V(p)\}$;
8. (**A-contrast$^+$**) if $\exists p \in \mathsf{At}\colon P \subseteq [\![\pm p]\!]$, then $\bigcup \mathbf{r}(P,w) \subseteq \overline{P}$;
9. (**fallibilism**) $\exists P \subseteq W\ \exists w \in W\ \exists A \in \mathbf{r}(P,w)\colon w \in P$ and $A \subsetneq \overline{P}$.

Proof. Take the set of worlds and valuation in Fig. 5.7 with the $\mathbf{r}$ function in Table 5.3, and one can check that all nine conditions of Proposition 5.4 are satisfied.[23] □

[23]Table 5.3 only specifies the $\mathbf{r}$ function for w_1, but for the other worlds we can simply use the infallibilist formula $\mathbf{r}(P,w_i) = \{\overline{P} \cap W_w\}$ and all of the postulates will be satisfied for the whole model. We could instead define $\mathbf{r}$ in an appropriate fallibilist way for each of the other words, but I leave that as an exercise to the reader. As observed in footnote 16 of §2.4, to plausibly model the agent's knowledge at worlds other than w_1, we should add more structure to the model.

Proposition 5.4 is even easier to see in the color-coded graphical form in Fig. 5.11. As before, each row in Fig. 5.11 displays a proposition, *the set of worlds outlined in blue*, along with the alternative set(s) for the proposition. In a given row, if two possibilities are shaded in the same color, then they belong to the same alternative set. Hence it is only row six that has two alternative sets, each containing one possibility. I have shaded w_2 in rows two and six the same color orange in order to indicate that the orange path to knowing $p \vee \neg s$ in row six is the path that goes via knowing p in row two, consistent with the **overlap** postulate. We can construct larger and more complicated models with many alternative sets for a given proposition, but we can already see from the simple model in Fig. 5.11 how the key ideas of the Multipath Picture from §5.1.1 and §5.1.2 allow us to combine no vacuous knowledge with single-premise logical closure. Moreover, the model in Fig. 5.11 delivers what I take to be exactly the right verdicts—from the perspective of a fallibilist who denies full closure against the skeptic—about what the agent knows. While Nozick's picture in Fig. 5.9 strangely shows the agent as knowing the strong $p \wedge \neg s$ without knowing the weaker $p \vee \neg s$, our new picture in Fig. 5.11 correctly reverses these verdicts.

To claim that we have solved the twin problems of Vacuous Knowledge and Containment, we should show not only that **noVK** and **cover** are consistent in the Multipath Picture, but also that the model witnessing their consistency is a natural one for fallibilists. Indeed, the model in Fig. 5.11 can been seen as generated from minimal fallibilist assumptions. Step 1: start with the *infallibilist* model in Fig. 5.10 where for all P, the only alternative set for P is $\overline{P}$; so $\mathbf{r}_1(P, w) = \{\overline{P}\}$. Step 2: since the minimal fallibilist assumption is that knowing p in w_1 (relative to ordinary contexts) does not require ruling out skeptical worlds, modify $\mathbf{r}_1$ to $\mathbf{r}_2$ by taking all s-worlds out of all alternative sets for $[\![p]\!]$; so $\mathbf{r}_2([\![p]\!], w_1) = \{\overline{[\![p]\!]} \cap \overline{[\![s]\!]}\}$. Step 3: since knowing p is a path to knowing propositions logically weaker than p, modify $\mathbf{r}_2$ to $\mathbf{r}_3$ by adding the alternative sets for $[\![p]\!]$ to the set of alternative sets for any weaker Q; so if $[\![p]\!] \subseteq Q$, then $\mathbf{r}_3(Q, w) = \mathbf{r}_2(Q, w) \cup \mathbf{r}_2([\![p]\!], w)$. The result is Fig. 5.11.

Not only does this solve the problems of Vacuous Knowledge and Containment, but also the justification given for our witnessing model raises a challenge for fallibilists who insist on full (multi-premise) closure: try to justify step-by-step your further

$$\mathsf{U}(w_1) = \{w_1, w_3\}$$
$$\mathsf{u}(P, w_1) = \mathsf{U}(w_1) \cap \overline{P}$$

$\mathbf{r}([\![p \wedge \neg p]\!], w_1) =$	$\mathbf{r}(\emptyset, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p \wedge s]\!], w_1) =$	$\mathbf{r}(\{w_4\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p \wedge \neg s]\!], w_1) =$	$\mathbf{r}(\{w_1\}, w_1)$	$= \{\{w_2, w_3, w_4\}\}$	$\mathfrak{M}, w_1 \not\models K(p \wedge \neg s)$
$\mathbf{r}([\![\neg p \wedge s]\!], w_1) =$	$\mathbf{r}(\{w_3\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![\neg p \wedge \neg s]\!], w_1) =$	$\mathbf{r}(\{w_2\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p]\!], w_1) =$	$\mathbf{r}(\{w_1, w_4\}, w_1)$	$= \{\{w_2\}\}$	$\mathfrak{M}, w_1 \models Kp$
$\mathbf{r}([\![s]\!], w_1) =$	$\mathbf{r}(\{w_3, w_4\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p \leftrightarrow s]\!], w_1) =$	$\mathbf{r}(\{w_2, w_4\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p \leftrightarrow \neg s]\!], w_1) =$	$\mathbf{r}(\{w_1, w_3\}, w_1)$	$= \{\{w_2, w_4\}\}$	$\mathfrak{M}, w_1 \models K(p \leftrightarrow \neg s)$
$\mathbf{r}([\![\neg s]\!], w_1) =$	$\mathbf{r}(\{w_1, w_2\}, w_1)$	$= \{\{w_3, w_4\}\}$	$\mathfrak{M}, w_1 \not\models K\neg s$
$\mathbf{r}([\![\neg p]\!], w_1) =$	$\mathbf{r}(\{w_2, w_3\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p \vee s]\!], w_1) =$	$\mathbf{r}(\{w_1, w_3, w_4\}, w_1)$	$= \{\{w_2\}\}$	$\mathfrak{M}, w_1 \models K(p \vee s)$
$\mathbf{r}([\![p \vee \neg s]\!], w_1) =$	$\mathbf{r}(\{w_1, w_2, w_4\}, w_1)$	$= \{\{w_2\}, \{w_3\}\}$	$\mathfrak{M}, w_1 \models K(p \vee \neg s)$
$\mathbf{r}([\![p \rightarrow s]\!], w_1) =$	$\mathbf{r}(\{w_2, w_3, w_4\}, w_1)$	$= \emptyset$	
$\mathbf{r}([\![p \rightarrow \neg s]\!], w_1) =$	$\mathbf{r}(\{w_1, w_2, w_3\}, w_1)$	$= \{\{w_4\}\}$	$\mathfrak{M}, w_1 \models K(p \rightarrow \neg s)$
$\mathbf{r}([\![p \vee \neg p]\!], w_1) =$	$\mathbf{r}(W, w_1)$	$= \{\emptyset\}$	$\mathfrak{M}, w_1 \models K(p \vee \neg p)$

Table 5.3: partial representation of an MA model for Proposition 5.4

modifications to the model in Fig. 5.11, which are required to restore full closure, on the basis of minimal fallibilist assumptions. In addition to raising this challenge, in the next section I will present impossibility results to the effect that fallibilists who insist on full closure will lead us back into the Problem of Vacuous Knowledge.

5.2.4 Finer-Grained Structure

In this section, I return to an issue first raised in §5.1.4, namely the distinction between logical structure and syntactic structure. As we have seen, by adding points to a model we can pull apart the extensions of logically inequivalent formulas, to which we can then assign different alternative sets in our MA model; but we cannot pull apart the extensions of logically equivalent formulas, so we cannot assign different alternative sets to them in our MA model. One might think that given the example

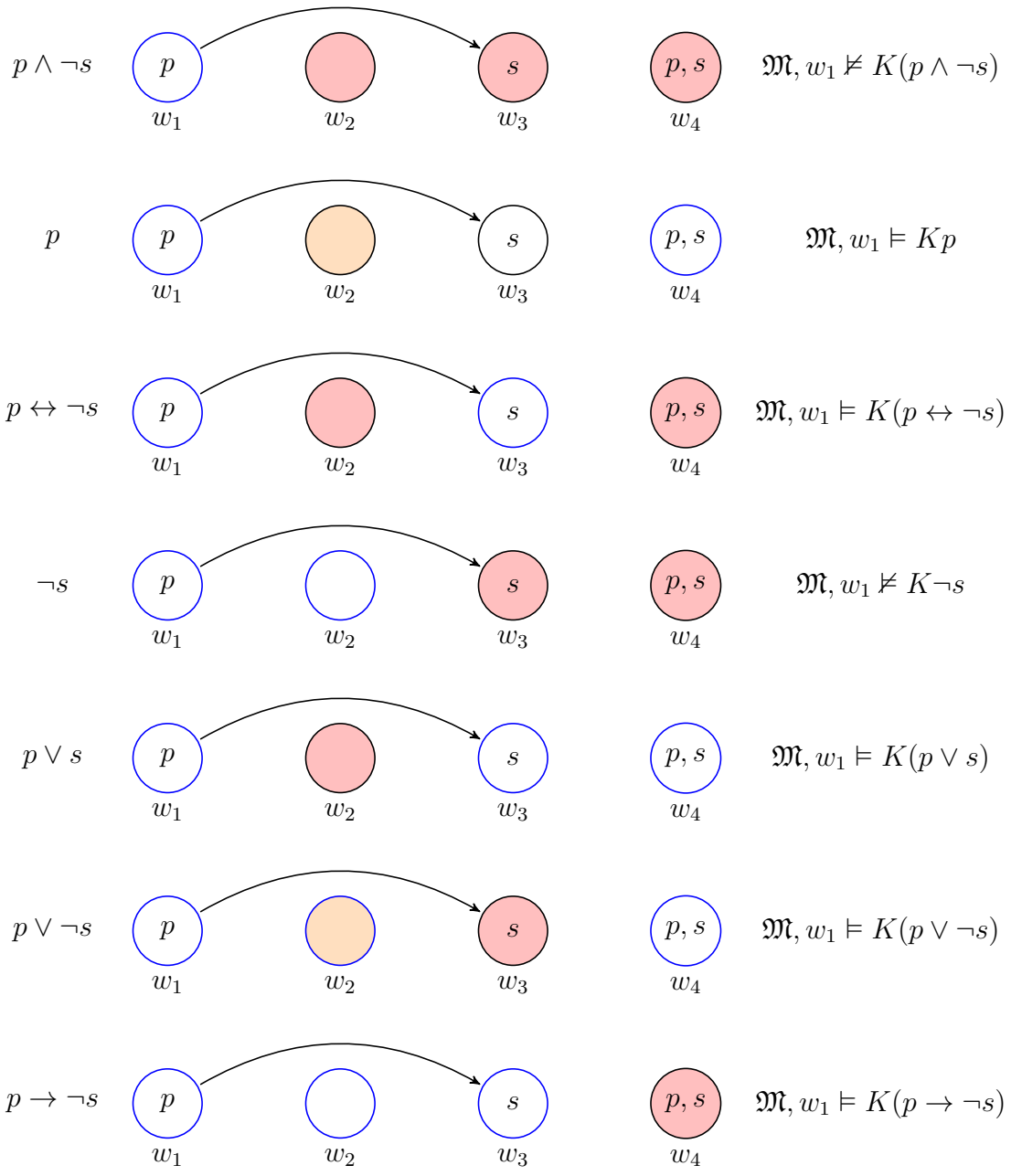

Figure 5.11: partial representation of an MA model for Proposition 5.4

of multiple paths to knowing a disjunction in §5.1.1, it is not clear why the logical equivalence of φ and ψ should imply that φ and ψ have the same alternative sets. For one formula that is not syntactically disjunctive may be logically equivalent to another formula that is: e.g., p is logically equivalent to $(p \wedge q) \vee (p \wedge \neg q)$. One might think that we should allow that in some cases, while there may be only one path to knowing p, there are multiple paths to knowing $(p \wedge q) \vee (p \wedge \neg q)$. To see where this idea leads, let us define a new class of models in which logically equivalent formulas may have different alternative sets. Form is the set of formulas of our language.

Definition 5.5 (SMA Model)**.** A *syntactic multipath alternatives* model is a tuple $\mathfrak{M}$ of the form $\langle \mathbf{W}, u, r, V \rangle$ where $\mathbf{W} = \langle W, \{W_w\}_{w \in W} \rangle$ with W a non-empty set and $w \in W_w \subseteq W$; $u\colon \mathsf{Form} \times W \to \mathcal{P}(W)$, $r\colon \mathsf{Form} \times W \to \mathcal{P}(\mathcal{P}(W))$, and $V\colon \mathsf{At} \to \mathcal{P}(W)$.

We define truth in a SMA model following the same pattern as Definition 5.2.

Definition 5.6 (Truth in a SMA Model)**.** Given a SMA model $\mathfrak{M} = \langle \mathbf{W}, u, r, V \rangle$ with $w \in W$ and a formula φ in the epistemic-alethic language, we define $\mathfrak{M}, w \vDash \varphi$ as follows (with propositional and $\Box$ cases as usual):

$$\mathfrak{M}, w \vDash K\varphi \quad \text{iff} \quad \exists A \in r(\varphi, w)\colon A \cap u(\varphi, w) = \emptyset.$$

I will now show that if single-premise logical closure holds, then the move from MA to SMA models is not necessary for the purposes of representing IALs' knowledge. In order for single-premise logical closure to hold, it must be that if φ is a logical consequence of ψ, then every path to knowing ψ covers a path to knowing φ:[24]

$$\forall B \in r(\psi, w)\, \exists A \in r(\varphi, w)\colon A \subseteq B. \tag{5.1}$$

Now suppose that φ and ψ are logically equivalent, so we also have

$$\forall B \in r(\varphi, w)\, \exists A \in r(\psi, w)\colon A \subseteq B. \tag{5.2}$$

Together (5.1) and (5.2) are consistent with $r(\varphi, w) \neq r(\psi, w)$. Hence we can combine

[24]Here I assume that u satisfies the $\mathrm{RO}_{\exists\forall}$ condition from §4.B.

single-premise logical closure with the idea that, e.g., p and $(p \wedge q) \vee (p \wedge \neg q)$ have different alternative sets. However, we can show that their having different alternative sets is a not a reflection of a real distinction, but rather a reflection of redundancy. To do so, we first define an operation on r functions that eliminates redundancies.

Definition 5.7 (r_- function)**.** Given a function $r\colon \mathsf{Form} \times W \to \mathcal{P}(\mathcal{P}(W))$, define the function $r_-\colon \mathsf{Form} \times W \to \mathcal{P}(\mathcal{P}(W))$ as follows:

$$r_-(\varphi, w) = \{B \in r(\varphi, w) \mid \forall A \in r(\varphi, w)\colon A \not\subseteq B\}.$$

The following fact states the sense in which the extra alternative sets in $r(\varphi, w)$ that are not in $r_-(\varphi, w)$ are redundant for representing the agent's knowledge.

Fact 5.1 (From r to r_-)**.** Where $\mathfrak{M} = \langle \mathbf{W}, u, r, V\rangle$ is an SMA model and $\mathfrak{M}_- = \langle \mathbf{W}, u, r_-, V\rangle$, the following holds for any formula α:

$$\mathfrak{M}, w \vDash \alpha \text{ iff } \mathfrak{M}_-, w \vDash \alpha.$$

The final fact in the argument is that while (5.1) and (5.2) are jointly consistent with $r(\varphi, w) \neq r(\psi, w)$, they are *not* jointly consistent with $r_-(\varphi, w) \neq r_-(\psi, w)$.

Fact 5.2. If (5.1) and (5.2) hold for r_-, then $r_-(\varphi, w) = r_-(\psi, w)$.

Proof. Assume (5.1) and (5.2) hold for r_- and suppose for *reductio* that there is some $C \in r_-(\varphi, w)$ such that $C \notin r_-(\psi, w)$. It follows by (5.2) that there is $B \in r_-(\psi, w)$ such that $B \subsetneq C$, so by (5.1) there is $A \in r_-(\varphi, w)$ such that $A \subseteq B$. But then there are $A, C \in r_-(\psi, w)$ such that $A \subsetneq C$, contradicting the definition of r_-. It follows that $r_-(\varphi, w) \subseteq r_-(\psi, w)$, and an analogous argument shows $r_-(\psi, w) \subseteq r_-(\varphi, w)$. □

Together Facts 5.1 and 5.2 show that if single-premise logical closure holds, then the *non-redundant* alternative set representations for logically equivalent formulas must be the same. In the *redundant* representation, $(p \wedge q) \vee (p \wedge \neg q)$ may indeed have multiple alternative sets while p has only one, but this is because there are alternative sets $A, B \in r((p \wedge q) \vee (p \wedge \neg q), w)$ such that $A \subsetneq B$, so B is redundant. In particular,

B may be the alternative set for the path to knowing $(p \wedge q) \vee (p \wedge \neg q)$ via knowing $(p \wedge q)$, while A may be a strictly smaller alternative set for the path to knowing the disjunction without knowing either disjunct. However, when we eliminate redundant alternative sets like B, we have $r_-(p, w) = r_-((p \wedge q) \vee (p \wedge \neg q), w)$.

If $r_-(\varphi, w) = r_-(\psi, w)$ whenever φ and ψ are logically equivalent, then we might as well use the $\mathbf{r}$ function that ignores syntax, setting $\mathbf{r}([\![\varphi]\!], w) = r_-(\varphi, w) = r_-(\psi, w) = \mathbf{r}([\![\psi]\!], w)$. One difference is that with the $\mathbf{r}$ function, if α and β are logically *in*equivalent formulas such that $[\![\alpha]\!] = [\![\beta]\!]$ happens to hold in our model, then we must add points to the model such that $[\![\alpha]\!] \neq [\![\beta]\!]$ in order to have $\mathbf{r}([\![\alpha]\!], w) \neq \mathbf{r}([\![\beta]\!], w)$. By contrast, with the r function, we can have both $[\![\alpha]\!] = [\![\beta]\!]$ and $r_-([\![\alpha]\!], w) \neq r_-([\![\beta]\!], w)$. Whether to use $\mathbf{r}$ or r becomes a matter of modeling preference: in MA models, we may have to add more points, whereas in SMA models, we have to define r on more inputs (Form vs. $\mathcal{P}(W)$). Either way, the importance of Facts 5.1 and 5.2 is to show that given closure under logical equivalence, the multiple paths to knowledge of φ arise from its logical structure, not its syntactic structure.

Where SMA models become essential is in representing views of knowledge that reject closure under logical equivalence even for IALs, views according to which there are logically equivalent φ and ψ such that knowing φ requires more empirical elimination of possibilities than knowing ψ. For example, as I will discuss in §6.2.2, Dretske [1970] seems to implicitly reject closure under logical equivalence while explicitly accepting closure principles like $K(\varphi \wedge \psi) \rightarrow K\varphi$ and $K\varphi \rightarrow K(\varphi \vee \psi)$. While we cannot represent such views in MA models, we can do so in SMA models. Not only can we represent such views, but we can solve a problem for them. Recall from the impossibility results of §4.B.1 that in the framework of Fallibilism 1.0, validating either of these principles (even without full single-premise closure) leads to the Problem of Vacuous Knowledge.[25] By contrast, by taking advantage of the multipath picture, we can transfer the consistency result of Proposition 5.4 from MA models to SMA models to show that fallibilists can accept *any* special cases of single-premise closure, such as $K(\varphi \wedge \psi) \rightarrow K\varphi$ and $K\varphi \rightarrow K(\varphi \vee \psi)$, while avoiding vacuous knowledge.

[25]To be precise, the impossibility result for $K(\varphi \wedge \psi) \rightarrow K\varphi$ in Proposition 4.4 also assumed a special case of closure under logical equivalence, $K\varphi \leftrightarrow K((\varphi \vee \psi) \wedge \varphi)$.

The following fact gives the simple transformation from MA models to SMA models.

Fact 5.3 (From MA to SMA Models)**.** Given a MA model $\mathfrak{M} = \langle \mathbf{W}, \mathsf{u}, \mathbf{r}, V \rangle$, define a SMA model $\mathfrak{N} = \langle \mathbf{W}, u, r, V \rangle$ by

$$r(\varphi, w) = \mathbf{r}(\llbracket \varphi \rrbracket, w); \tag{5.3}$$

$$u(\varphi, w) = \mathsf{u}(\llbracket \varphi \rrbracket, w). \tag{5.4}$$

For all $w \in W$ and formulas φ, $\mathfrak{M}, w \vDash \varphi$ iff $\mathfrak{N}, w \vDash \varphi$, and if $\mathfrak{M}$ satisfies any of the conditions of Proposition 5.4, then $\mathfrak{N}$ satisfies the corresponding conditions:

1. (*r-RofA*) if $w \in \overline{\llbracket \varphi \rrbracket}$, then $\mathbf{r}(\varphi, w) = \emptyset$;
2. (*enough*) if $w \in \llbracket \varphi \rrbracket$, then $\exists A \in \mathbf{r}(\varphi, w)\colon A \subseteq \overline{\llbracket \varphi \rrbracket}$;
3. (*noVK*) if $\llbracket \varphi \rrbracket \neq W_w$, then $\emptyset \notin \mathbf{r}(\varphi, w)$;
4. (*overlap*) $\forall A \in \mathbf{r}(\varphi, w)$: if $A \cap \llbracket \varphi \rrbracket \neq \emptyset$, then $\exists \psi\colon \llbracket \psi \rrbracket \subsetneq \llbracket \varphi \rrbracket$ and $A \in \mathbf{r}(\psi, w)$;
5. (*cover*) if $\llbracket \psi \rrbracket \subseteq \llbracket \varphi \rrbracket$, then $\forall B \in \mathbf{r}(\psi, w)\ \exists A \in \mathbf{r}(\varphi, w)\colon A \subseteq B$;
6. (*r-possible*) $\bigcup \mathbf{r}(\varphi, w) \subseteq W_w$;
7. (*A-space*) $\forall \mathsf{S} \subseteq \mathsf{At}\ \exists s \in W\colon \mathsf{S} = \{p \in \mathsf{At} \mid s \in V(p)\}$;
8. (*A-contrast*$^+$) if $\exists p \in \mathsf{At}\colon \llbracket \varphi \rrbracket \subseteq \llbracket \pm p \rrbracket$, then $\bigcup \mathbf{r}(\varphi, w) \subseteq \overline{\llbracket \varphi \rrbracket}$;
9. (*fallibilism*) $\exists \varphi\, \exists w \in W\, \exists A \in \mathbf{r}(\varphi, w)\colon w \in \llbracket \varphi \rrbracket$ and $A \subsetneq \overline{\llbracket \varphi \rrbracket}$.

Together Proposition 5.4 and Fact 5.3 show that there are SMA models satisfying conditions 1 - 9. Of course, it follows that any weakenings of conditions 1 - 9 are also consistent. The important point, for our purposes, is that the multipath picture allows fallibilists who reject full single-premise closure to accept any weakenings of *cover* for specific instances of single-premise closure, while avoiding vacuous knowledge.

I will return to views like Dretske's that weaken single-premise closure in §6.2.2. First, however, we must consider what happens if we go in the other direction and add the assumptions necessary for full multi-premise closure in the multipath picture.

5.3 Full Closure

By Proposition 5.2, full epistemic closure requires the **combine** assumption on $\mathbf{r}$. What do I say to fallibilists who wish to add **combine** to the Five Postulates? I cannot say that the addition will result in inconsistency, since it will not.

Proposition 5.5 (Consistency II)**.** The nine conditions of Proposition 5.4 plus **combine** are consistent.

Proof. The model in Fig. 5.12 witnesses the consistency of the ten conditions. □

The model in Fig. 5.12 can be seen as coming from my preferred model in Fig. 5.11 by two additional steps. Step 4: if we assume that ruling out w_3 is not necessary in order to know $p \wedge \neg s$ or to know $\neg s$, then we modify $\mathbf{r}_3$ to $\mathbf{r}_4$ by taking w_3 out of all alternative sets for $[\![p \wedge \neg s]\!]$ and $[\![\neg s]\!]$; so $\mathbf{r}_4([\![p \wedge \neg s]\!], w_1) = \{\overline{[\![p \wedge \neg s]\!]} \setminus \{w_3\}\}$ and $\mathbf{r}_4([\![\neg s]\!], w_1) = \{\overline{[\![\neg s]\!]} \setminus \{w_3\}\}$. Step 5: since knowing $\neg s$ is a path to knowing propositions logically weaker than $\neg s$, modify $\mathbf{r}_4$ to $\mathbf{r}_5$ by adding the alternative sets for $[\![\neg s]\!]$ to the set of alternative sets for any weaker Q; so if $[\![\neg s]\!] \subseteq Q$, then $\mathbf{r}_5(Q, w) = \mathbf{r}_4(Q, w) \cup \mathbf{r}_4([\![\neg s]\!], w)$. We can do the same for $p \wedge \neg s$, but recall from the remark after Definition 5.2 that we display our models in a non-redundant form so that if $A, B \in \mathbf{r}(P, w)$ and $A \subsetneq B$, then we do not display B as an alternative set. The result is the model in Fig. 5.12, and observe that it satisfies **combine**.

Although as a technical matter the model in Fig. 5.12 establishes Proposition 5.5, as a conceptual matter there are two problems. First, what is the justification for Step 4? In my view, it is not a minimal fallibilist assumption that knowing $p \wedge \neg s$ and knowing $\neg s$ do not require ruling out the world w_3 where s is true. Its removal from the alternative sets seems to be an *ad hoc* effort to enforce closure. Moreover, if knowing $p \wedge \neg s$ and $\neg s$ does not require ruling out w_3, why should it require ruling out the even more bizarre (perhaps impossible) world w_4 in which p and s are both true? I see no good argument for this position, but note that if we were to remove w_4 from the alternative sets, then we would violate **noVK** in row four.

The last observation suggests a problem lurking for full closure. So far I have been assuming that the mundane proposition p and the skeptical hypothesis s are

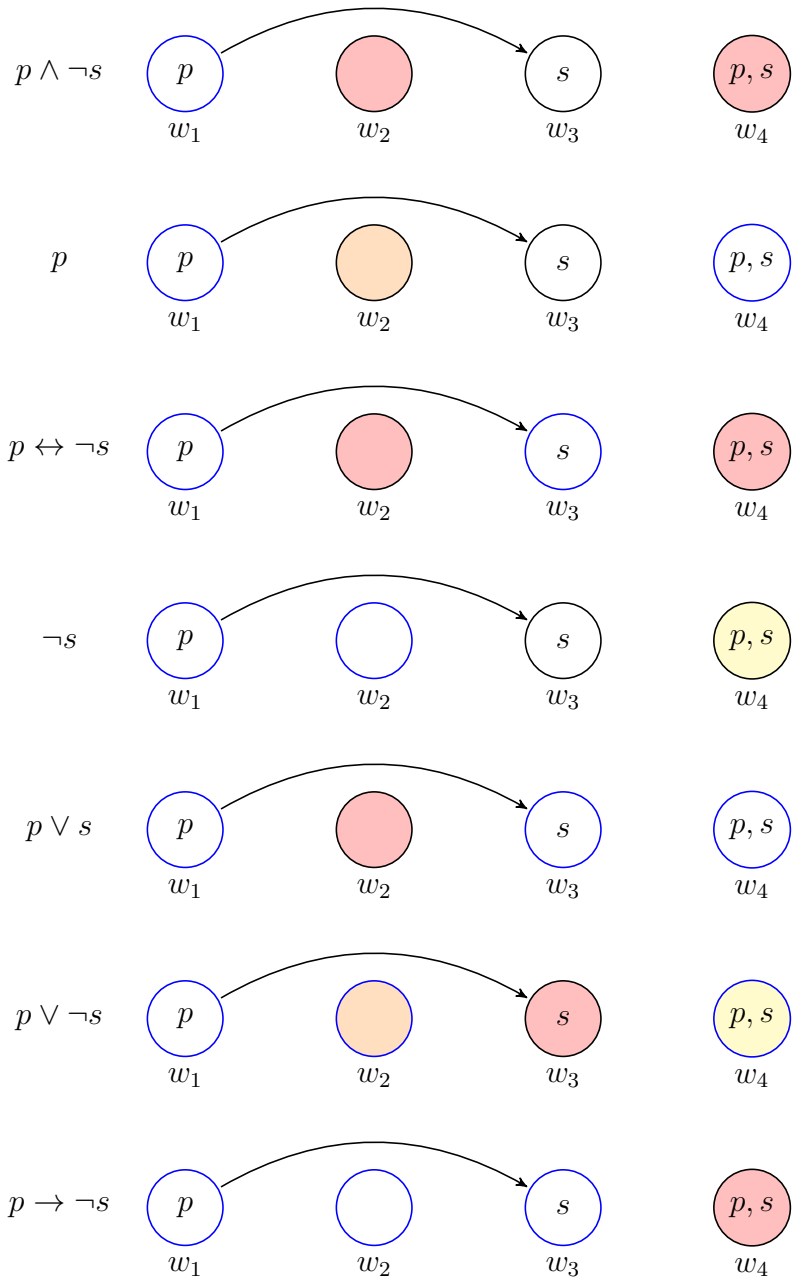

Figure 5.12: partial representation of an MA model for Proposition 5.5

incompatible in some weaker sense than $p \to \neg s$ being metaphysically necessary. But let us now assume that $p \to \neg s$ is metaphysically necessary at the actual world w_1, so any $(p \wedge s)$-points like w_4 are not possible worlds in W_{w_1}. If we also adopt **r-possible** from §5.1.3, so alternative sets for propositions only contain possible worlds, then we must remove w_4 from the alternative sets. Doing so starting from my preferred model in Fig. 5.11 and making adjustments to satisfy **cover** and **enough**, we obtain the model in Fig. 5.13 that still satisfies all of the conditions in Proposition 5.4. However, if we remove w_4 from the alternative sets in the "closed" model in Fig. 5.12, then we violate **noVK** in row four. In fact, the problem with full closure here is general, not merely with the model in Fig. 5.12. For the following proposition, think of P as the mundane proposition and S as the metaphysically incompatible skeptical hypothesis. After giving the proof, I will explain the argument informally below.

Proposition 5.6 (Impossibility III)**.** There is no MA model satisfying the following conditions (but there are MA models satisfying all but **combine**, as in Fig. 5.13):

1. (**SK**) For some $w \in W$ and $P, S \subseteq W$:
   (a) $w \in \overline{S}$;
   Read: the skeptical hypothesis S is false at w.
   (b) $P \cap S \cap W_w = \emptyset$;
   Read: P and S are incompatible as a matter of metaphysical necessity.
   (c) $\forall A \in \mathbf{r}(P, w)$: $A \cap S = \emptyset$;
   Read: knowing P does not require ruling out skeptical S-worlds.
   (d) $\forall A \in \mathbf{r}(\overline{S}, w)$: $A \cap S \neq \emptyset$;
   Read: knowing not-S requires ruling out some S-worlds.
   Note: if $S = [\![s]\!]$ for $s \in \mathsf{At}$, then (d) follows from **noVK** and **A-contrast**.
2. (**r-possible**) $\bigcup \mathbf{r}(P, w) \subseteq W_w$;
3. (**enough**) if $w \in P$, then $\exists A \in \mathbf{r}(P, w)$: $A \subseteq \overline{P}$;
4. (**cover**) if $Q \subseteq P$, then $\forall B \in \mathbf{r}(Q, w)\ \exists A \in \mathbf{r}(P, w)$: $A \subseteq B$;

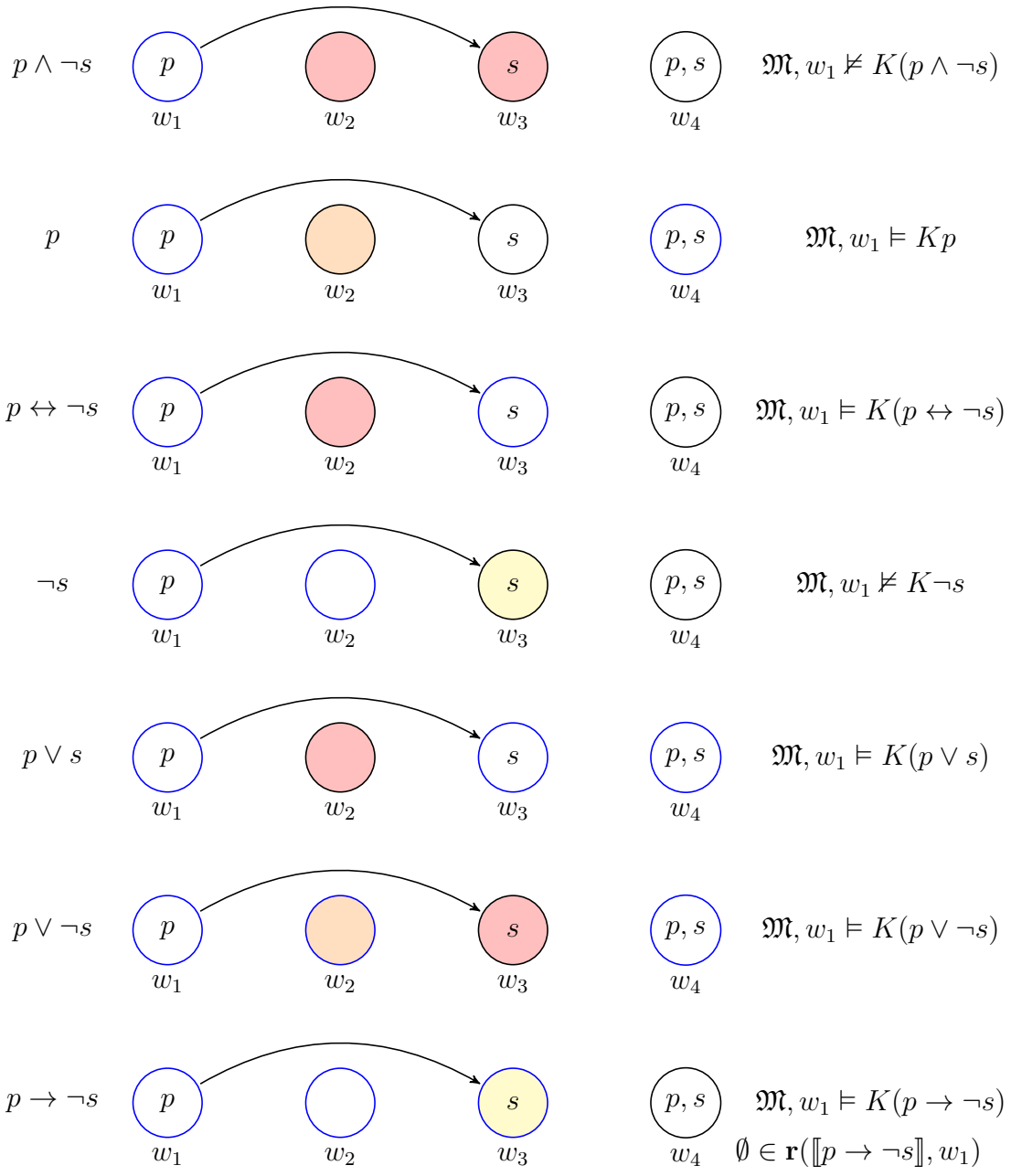

Figure 5.13: partial representation of an MA model for Proposition 5.4 ($w_4 \notin W_{w_1}$)

5. (**combine**)

$$\forall P, P' \subseteq W\ \forall A \in \mathbf{r}(P, w)\ \forall A' \in \mathbf{r}(P', w)\ \exists B \in \mathbf{r}(P \cap P', w)\colon B \subseteq A \cup A'.$$

Proof. By **cover**, for all $B \in \mathbf{r}(P \cap \overline{S})$, there is some $A \in \mathbf{r}(\overline{S}, w)$ such that $A \subseteq B$. It follows by (d) that

$$\forall B \in \mathbf{r}(P \cap \overline{S})\colon B \cap S \neq \emptyset. \tag{5.5}$$

By (a) and **enough**, there is some $A \in \mathbf{r}(\overline{P \cap S}, w)$ such that $A \subseteq P \cap S$, which with (b) and **r-possible** implies $A = \emptyset$, so

$$\emptyset \in \mathbf{r}(\overline{P \cap S}, w). \tag{5.6}$$

Given $P \cap (\overline{P \cap S}) = P \cap \overline{S}$, (5.6) and **combine** together imply

$$\forall A \in \mathbf{r}(P, w)\ \exists B \in \mathbf{r}(P \cap \overline{S}, w)\colon B \subseteq A \cup \emptyset. \tag{5.7}$$

Observe that (5.7), (c), and (5.5) are inconsistent. □

Remark 5.4. Here is the argument for Proposition 5.6 informally. Assume $p \to \neg s$ is necessary and knowable *a priori*, without the empirical elimination of any possibilities, as Hawthorne [2004a, 39] assumes for some skeptical hypotheses. Add to this the minimal fallibilist assumption that knowing p (relative to ordinary contexts) does not require ruling out any s-worlds. We must ask: does knowing $p \wedge \neg s$ require ruling out any s-worlds? Case 1: suppose the answer is 'no'. Then assuming that knowing $p \wedge \neg s$ suffices for knowing $\neg s$, it follows that knowing $\neg s$ does not require ruling out any s-worlds, which violates the conjunction of **noVK** and **A-contrast**. Case 2: suppose the answer is 'yes'. Then the agent can know p and $p \to \neg s$ individually, since this does not require ruling out s-worlds, without knowing something logically equivalent to the *conjunction* of p and $p \to \neg s$, namely $p \wedge \neg s$, which violates multi-premise closure by violating **combine**. In short, full closure requires vacuous knowledge.

I have already explained in Chapter 4 why I reject vacuous knowledge, so I conclude as in Claim 5.2 that fallibilists should reject **combine** and multi-premise closure.

Hence we are lead to reject the principle $(K\varphi \wedge K\psi) \rightarrow K(\varphi \wedge \psi)$ for reasons distinct from the typical ones concerning probability and "accumulation of risk" (see, e.g., Skyrms 1967; Goldman 1975, Hawthorne 2004a, §1.6).[26,27] Those who can stomach the idea that an agent may know p and know $p \rightarrow \neg s$ without knowing $\neg s$ may pause at the idea that such an agent could fail to know $p \wedge (p \rightarrow \neg s)$. But one should not be fooled by syntax into thinking that there is something harmless about putting p and $(p \rightarrow \neg s)$ together with $\wedge$. We must remind ourselves that $p \wedge (p \rightarrow \neg s)$ is logically stronger than $\neg s$, so if it is difficult to know the latter, then it must also be difficult to know the former. The sense that it should be easy to know $p \wedge (p \rightarrow \neg s)$ (just put the $\wedge$ in between!) is a cognitive illusion induced by too much focus on syntax.

Almost the same argument as that of Remark 5.4 applies against closure under known strict bi-implication (recall §5.1.5). Assume that $p \leftrightarrow (p \wedge \neg s)$ is necessary and knowable *a priori*, without the empirical elimination of possibilities, as Hawthorne [2004a, 39] assumes for some skeptical hypotheses. Add to this the minimal fallibilist assumption that knowing p (relative to ordinary contexts) does not require ruling out any s-worlds. We must ask: does knowing $p \wedge \neg s$ require ruling out any s-worlds? Case 1: suppose the answer is 'no'. Then assuming that knowing $p \wedge \neg s$ suffices for knowing $\neg s$, it follows that knowing $\neg s$ does not require ruling out any s-worlds, which

[26]Assuming (incorrectly) that the only worries about multi-premise closure have to do with accumulation of risk, Hawthorne [2004a, 35n88] writes:

> [T]hose who accept SPC [single-premise closure] will also, presumably, be happy with the following special case of MPC [multi-premise closure]: Necessarily, if one knows p and, in conjunction with a set of premises that are known with certainty a priori, competently deduces q, thereby coming to believe q, retaining one's knowledge of p and of that set of premises throughout, then one comes to know q. The standard kind of worry concerning MPC—that small risks add up to big risks—has far less force with regard to this special case.

As explained in Remark 5.4, I reject this special case of multi-premise closure for different reasons.

[27]Lasonen-Aarnio [2008] argues that single- and multi-premise closure "come as a package: either both will have to be rejected or both will have to be revised" (157). The argument for this claim is that both single- and multi-premise closure are susceptible to similar kinds of "accumulation of risk" failures, given the fallibility of humans in deductive reasoning. There are several reasons why this is orthogonal to our discussion. For one thing, as discussed in §2.1, I have been considering which closure principles hold for ideally astute logicians (IALs) who are deductively infallible. Setting this point aside, my claim is that single-premise logical closure does not lead to the Problem of Vacuous Knowledge, while multi-premise logical closure does. This claim is consistent with Lasonen-Aarnio's view that both types of closure are problematic for accumulation of risk reasons.

violates the conjunction of **noVK** and **A-contrast**. Case 2: suppose the answer is 'yes'. Then the agent can know p, since this does not require ruling out s-worlds, without knowing $p \wedge \neg s$, which violates closure under known strict bi-implication. In short, closure under known strict bi-implication requires either vacuous knowledge or a rejection of $K(p \wedge \neg s) \rightarrow K\neg s$. In §6.1.2, I will argue against $Kp \leftrightarrow K(p \wedge \neg s)$.

The formal version of the result can be stated as in Proposition 5.7, using the fact from Proposition 5.2 that closure under strict bi-implication corresponds to the **M-equiv** condition. One can easily prove a similar result for closure under *known* strict bi-implication, but the corresponding condition on $\mathbf{r}$ is more complicated (and involves the u function), so I omit it. The informal argument above suffices.

Proposition 5.7 (Impossibility IV)**.** There is no MA model satisfying the following:

1. (**SK**) For some $w \in W$ and $P, S \subseteq W$:
   (a) $P \cap S \cap W_w = \emptyset$;
   (b) $\forall A \in \mathbf{r}(P, w)$: $A \cap S = \emptyset$;
   (c) $\forall A \in \mathbf{r}(\overline{S}, w)$: $A \cap S \neq \emptyset$;
   Note: if $S = [\![s]\!]$ for $s \in \mathsf{At}$, then (d) follows from **noVK** and **A-contrast**.
2. (**M-equiv**) if $Q \cap W_w = P \cap W_w$, then $\mathbf{r}(Q, w) = \mathbf{r}(P, w)$;
3. (**cover**) if $Q \subseteq P$, then $\forall B \in \mathbf{r}(Q, w)\ \exists A \in \mathbf{r}(P, w)$: $A \subseteq B$.

Proof. By (a) and **M-equiv**, $\mathbf{r}(P, w) = \mathbf{r}(P \cap \overline{S}, w)$. By **cover**, for all $B \in \mathbf{r}(P \cap \overline{S})$ there is some $A \in \mathbf{r}(\overline{S}, w)$ such that $A \subseteq B$. It follows by (c) that for all $B \in \mathbf{r}(P \cap \overline{S})$, $B \cap S \neq \emptyset$, which contradicts (b) given $\mathbf{r}(P, w) = \mathbf{r}(P \cap \overline{S}, w)$. □

Using the SMA models of §5.2.4, we can state an analogue of Proposition 5.7 that shows the inconsistency using only the condition on r corresponding to $K(\varphi \wedge \psi) \rightarrow K\psi$, instead of full-single premise logical closure, but I leave this to the reader.

Some fallibilists who were willing to give up multi-premise closure to avoid vacuous knowledge may claim that it is worth biting the bullet on vacuous knowledge after all to keep closure under known strict bi-implication (or bi-implication known *a priori*).

In Chapter 6, I will further defend my choice to reject this closure principle rather than accept vacuous knowledge. For now, observe that if we run my three step argument at the end of §5.2.3 for generating the model in Fig. 5.11 from minimal fallibilist assumptions, together with the assumption that w_4 is an impossible point that should not be included in alternative sets, then we uniquely obtain the model in Fig. 5.13 that violates the closure principle. In this sense, minimal fallibilism leads directly to a rejection of closure under known strict bi-implication.

There are other ways of constructing MA models that satisfy the conditions of Proposition 5.4, besides the simple three step argument. In particular, we can construct such MA models from the world-orderings in RA models by a recursive construction, but I will not go into the details here.[28]

Let us take stock. Working with the Multipath Picture of Knowledge, we have developed a fallibilist theory of the $\mathbf{r}$ function that solves the problems of Vacuous Knowledge (§4.1) and Containment (§4.2), as shown by Proposition 5.4, without resorting to Knowledge Inflation (§4.3). To underscore the last point, in the next section I will present an account of the epistemic effect of "putting two and two together" that shows how competent deduction can extend knowledge without inflating it.

5.4 The Transfer Picture of Deduction

So far, we have studied the issue of closure *statically*, taking as our agent an ideally astute logician who has already finished deducing whatever follows from what she knows. Our job has been to reason what about what this "finished" agent knows. If you tell me that she knows $\varphi \wedge \psi$, may I conclude that she knows φ? In my view, yes. If you tell me that she knows φ and $\varphi \rightarrow \psi$, may I conclude that she knows ψ? In my view, not necessarily—I need to hear more about the case. And so on.

But what about the rest of us, who are never finished deducing whatever follows from what we know? In this final section, I will present a picture of the epistemic dynamics of deduction, or of putting two and two together, for us "unfinished" agents. I call this picture the Transfer Picture of Deduction. It is not a picture of what must

[28]Note added in ILLC version: see the Multipath Theorem of Holliday 2013b, §3.5.

go on in an agent's head in order for her to count as having competently deduced something. Instead, it is a picture of the *epistemic effect* of a competent deduction, compatible with different views about what it takes to competently deduce.

Let me digress momentarily to note that according to Harman and Sherman [2004, 495], talking as if there is an activity of *deducing* is a serious error:

> A more basic worry about the passage from Williamson is its presupposition that deduction is a kind of inference, something one does. Hawthorne apparently presupposes the same thing ....
>
> Surely, this confuses questions of implication with questions of inference. A deduction is a structured object, an abstract argument or proof. True, in order to check or exhibit implications, we sometimes construct arguments. And inference can be involved in that construction. But a deduction is the abstract argument that is constructed. Although constructing the argument is something someone does, the deduction itself is not something someone does. The deduction is not the constructing of the deduction .... The conclusion of a deduction is not in general the conclusion of an inference. (The conclusion of an inference might be that a certain construction is indeed a valid deduction. The whole argument is then the conclusion of the inference.)

I agree with Harman and Sherman that there is a sense of the *noun* 'deduction' referring to a structured, abstract object. But why can we (and Williamson and Hawthorne) not use the English *verb* 'deduce' to refer to what Sherlock Holmes does when he puts two and two together to conclude that so-and-so is the murderer? In the Oxford English Dictionary [2012], Definition 6.a. of 'deduction' is "the process of deducing or drawing a conclusion from a principle already known or assumed." I trust that in what follows, readers are capable of using context to determine when we are talking about the process and when we are talking about the abstract object.

Suppose that we have a model $\mathfrak{M}$ representing the epistemic state of our agent at some initial time. Subsequently, the agent puts two and two together and draws a conclusion from something(s) she already knows. If she gains new knowledge in this

way, then we need to update $\mathfrak{M}$ to $\mathfrak{M}'$ to reflect her new and improved epistemic state. In my view, this is a matter of updating the u function, which tells us which possibilities are uneliminated as alternatives for which propositions. Recall from §4.3 the following rule for updating the u function, inspired by Klein [1995]:

> Klein's Rule: if $\mathsf{r}(P,w) \cap \mathsf{u}(P,w) = \emptyset$ and the agent competently deduces Q from P, then update u to u' such that $\mathsf{r}(Q,w) \cap \mathsf{u}'(Q,w) = \emptyset$.[29]

In §4.3 I argued that this rule leads to a serious Problem of Knowledge Inflation.

My alternative view is based on the idea, familiar from the $\mathsf{RO}_{\forall\exists}$ theories of Chapter 3, that a single not-P-and-not-Q possibility x may be eliminated as an alternative for P but not eliminated as an alternative for Q. For example, suppose a scientist is trying to determine whether a hypothesis P is true. She draws up a list of various ways $x, y, z, \ldots$ in which the hypotheses could be false and starts running experiments to try to rule them out. When she rules out one of the ways x in which P could be false, she has eliminated x as an alternative for P. However, I would not say that she has thereby eliminated x as an alternative for every other proposition Q falsified by x, because for many of these Q, her inquiry concerning the question of P in which she eliminates x does not at all concern the question of Q. (One could have various theories about what it is for an inquiry to concern a question, but we can all think of clear cases in which a question was not at all a concern of one of our inquiries.) Now suppose that the scientist rules out the other possibilities and thereby comes to know P. Some time later, in a discussion at a conference on a different topic, someone asks her about her view on some proposition Q. She thinks to herself for a moment, when she realizes that Q follows from P. Assuming that she still knows P,[30] even if all of the details of her experiments concerning P are not fresh in her mind, what happens when she deduces Q from P? Here is my answer:

> Transfer: if $\exists A \in \mathbf{r}(P,w)\colon A \cap \mathsf{u}(P,w) = \emptyset$ and the agent competently deduces Q from P, then update u to $\mathsf{u}_{P\Rightarrow Q}$ such that $\mathsf{u}_{P\Rightarrow Q}(Q,w) = \mathsf{u}(Q,w) \cap \mathsf{u}(P,w)$.

[29] As noted before, this does not define a unique update rule, but picks out a class of update rules.
[30] Assume that realizing the consequence Q of P does not lead her to give up her belief in P.

In other words, all possibilities that were eliminated as alternatives for the known premise P *become eliminated as alternatives for the conclusion* Q. The epistemic effect of deduction is to transfer the elimination relation from alternatives for the premise to alternatives for the conclusion. Assuming full closure, the alternatives for the premise will cover the alternatives for the conclusion, so the transfer will correspond to our scientist coming to know Q from P. Deduction saves her the effort of doing new experiments to eliminate those alternatives for the conclusion or of having to recall the details of her past experiments to see how they bear on the conclusion. But deduction does *not* eliminate alternatives for the conclusion that were never eliminated as alternatives for the premise—that would be *knowledge inflation*.

Until this section, I have developed my framework in such a way as to be compatible with a variety of different views about what it is to "eliminate" a possibility. But the Transfer Picture of Deduction constrains the possible views of elimination. When the scientist trying to determine whether P is true draws up a list of various ways $x, y, z, \ldots$ in which P could be false and then runs an experiment whose result is incompatible with x,[31] this is a paradigm case of what might be called the "direct" empirical elimination of x as an alternative for P. According to the Transfer Picture, whether this counts as the direct, empirical elimination of x as an alternative for some Q depends on whether the inquiry that the experiment is part of concerns the question of Q. Moreover, according to the Transfer Picture, there is another way that the scientist could eliminate x as an alternative for some Q, other than this kind of direct, empirical elimination of x as an alternative for Q; namely, the agent could deduce Q from some known proposition P for which x has been eliminated as an alternative, either by direct, empirical elimination or by another deduction.

My proposal will actually be more general than Transfer. First, I will include the case of an agent deducing a conclusion from multiple premises. Second, I will modify

[31] A skeptic might say: have you really ruled out x? What if a demon is manipulating your measurement apparatus? But if x stands for a possibility in which the measurement apparatus is working correctly and the result of the measurement is 0, then in the actual world when the measurement apparatus is working correctly and the result of the measurement is 1, I would say that the scientist has ruled out x. One may want to add that the scientist knows that her apparatus is working properly, but as fallibilists we will say that she knows this provided that she has ruled out the relevant failure modes for the apparatus, which does not include the skeptic's demon.

the u function to allow that the agent may know φ and yet not know a logically equivalent ψ until she deduces ψ from φ. Before giving the definitions, I should note that this approach to the dynamics of deduction also works for the SA models of Chapter 3, but I will present it here in terms of the Multipath Picture of §5.2.

First, I will expand the epistemic language with new dynamic deduction operators of the form $\langle \varphi_1, \ldots, \varphi_n \Rightarrow \psi \rangle$, reading $\langle \varphi_1, \ldots, \varphi_n \Rightarrow \psi \rangle \chi$ as "after the agent competently deduces conclusion ψ from premises $\varphi_1, \ldots, \varphi_n$, χ is the case." Various ways of adding dynamic operators to the epistemic language to stand for acts of deduction or inference have been explored in depth in Velázquez-Quesada 2009, van Benthem and Velázquez-Quesada 2010, and Velázquez-Quesada 2011, but in a different semantic framework. I leave it to future work to compare these approaches to the one here.

Definition 5.8 (Deductive-Epistemic Language)**.** For a set of atomic sentences At, the deductive-epistemic language is generated by the following grammar:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid K\varphi \mid \langle \varphi, \ldots, \varphi \Rightarrow \varphi \rangle \varphi,$$

where $p \in \mathsf{At}$. Let Form be the set of all formulas.

To allow that an agent may know φ and yet not know a logically equivalent ψ until she deduces ψ from φ, I will now view the objects of an agent's knowledge not as sets-of-worlds propositions, but rather as structured propositions that can be distinguished as finely as formulas of our language. In particular, I will replace the function $\mathsf{u} \colon \mathcal{P}(W) \times W \to \mathcal{P}(W)$ with a function $\mathrm{u} \colon \mathsf{Form} \times W \to \mathcal{P}(W)$, so that a possibility may be eliminated as an alternative for one structured proposition/formula but not as an alternative for another, for the same reasons as discussed above.

However, the function $\mathbf{r} \colon \mathcal{P}(W) \times W \to \mathcal{P}(W)$ will be the same as before, reflecting my view that the *alternative sets* for a propositions are a function of the proposition's extension in logical space. The reason an agent may know φ and yet not know a logically equivalent ψ is not because knowing one of them requires empirically ruling out more or different possibilities than the other, but rather because an agent may rule out possibilities as alternatives for one of them without realizing the connection between φ and ψ. In short, it is a matter of the u function, not the $\mathbf{r}$ function.

Definition 5.9 (MDA Models)**.** A *multipath deductive alternatives* model is a tuple $\mathfrak{M} = \langle \mathbf{W}, \mathrm{u}, \mathbf{r}, V \rangle$ where $\mathbf{W}$, $\mathbf{r}$, and V are as in a MA model (Definition 5.1) and $\mathrm{u}\colon \mathsf{Form} \times W \to \mathcal{P}(W)$.

Hence MA models are MDA models where $[\![\varphi]\!] = [\![\psi]\!]$ implies $\mathrm{u}(\varphi, w) = \mathrm{u}(\psi, w)$. For now, I do not assume anything about the relation of $\mathrm{u}(\varphi, w)$ and $\mathrm{u}(\psi, w)$ for distinct φ and ψ (see after Proposition 5.9), although one may assume $\mathrm{u}(\varphi, w) \subseteq \overline{[\![\varphi]\!]}$.

Having moved from MA to MDA models, I can now state in general the rule for updating a MDA model $\mathfrak{M}$ to a new MDA model $\mathfrak{M}_{\varphi_1,\ldots,\varphi_n \Rightarrow \psi}$, reflecting the change in our agent's epistemic state as a result of her competently deducing ψ from $\varphi_1, \ldots, \varphi_n$. Of course, since an agent cannot deduce ψ from just *any* premises $\varphi_1, \ldots, \varphi_n$, we will need to put a further *precondition* on this action, but that will come later.

Definition 5.10 (Deductive Change)**.** Given a MDA model $\mathfrak{M} = \langle \mathbf{W}, \mathrm{u}, \mathbf{r}, V \rangle$, we define the model $\mathfrak{M}_{\varphi_1,\ldots,\varphi_n \Rightarrow \psi} = \langle \mathbf{W}, \mathrm{u}_{\varphi_1,\ldots,\varphi_n \Rightarrow \psi}, \mathbf{r}, V \rangle$ as follows. First, let

$$I_w = \{i \leq n \mid \mathfrak{M}, w \vDash K\varphi_i\}$$

be the set of indices of premises known at w. Then where $\chi \neq \psi$, for all $w \in W$, let

$$\begin{array}{rcl} \mathrm{u}_{\varphi_1,\ldots,\varphi_n \Rightarrow \psi}(\chi, w) & = & \mathrm{u}(\chi, w); \\ \mathrm{u}_{\varphi_1,\ldots,\varphi_n \Rightarrow \psi}(\psi, w) & = & \mathrm{u}(\psi, w) \cap \bigcap\limits_{i \in I_w} \mathrm{u}(\varphi_i, w). \end{array}$$

The idea is the same as before: all possibilities that were eliminated as alternatives for the known premises *become eliminated as alternatives for the conclusion.*

The range of worlds w for which we change $\mathrm{u}(\psi, w)$ matters for the agent's higher-order knowledge after the deduction. But since my main interest here is in uniterated knowledge, I set aside this subtlety and simply modify $\mathrm{u}(\psi, w)$ for all $w \in W$.

We are now ready to define truth for the deductive-epistemic language, taking the precondition formula $\mathsf{pre}(\varphi_1, \ldots, \varphi_n \Rightarrow \psi)$ for deduction as a parameter.

Definition 5.11 (Truth in MDA Models)**.** Given a MDA model $\mathfrak{M} = \langle \mathbf{W}, \mathrm{u}, \mathbf{r}, V \rangle$ with $w \in W$ and a formula φ in the deductive-epistemic language, we define $\mathfrak{M}, w \vDash \varphi$

as follows (with propositional cases as usual):

$$\begin{array}{lll}
\mathfrak{M}, w \vDash \Box\varphi & \text{iff} & W_w \subseteq [\![\varphi]\!]^{\mathfrak{M}}; \\
\mathfrak{M}, w \vDash K\varphi & \text{iff} & \exists A \in \mathbf{r}([\![\varphi]\!]^{\mathfrak{M}}, w)\colon A \cap \mathsf{u}(\varphi, w) = \emptyset; \\
\mathfrak{M}, w \vDash \langle \varphi_1, \ldots, \varphi_n \Rightarrow \psi \rangle \chi & \text{iff} & \mathfrak{M}, w \vDash \mathsf{pre}(\varphi_1, \ldots, \varphi_n \Rightarrow \psi) \text{ and} \\
 & & \mathfrak{M}_{\varphi_1, \ldots, \varphi_n \Rightarrow \psi}, w \vDash \chi.
\end{array}$$

One may consider various candidates for $\mathsf{pre}(\varphi_1, \ldots, \varphi_n \Rightarrow \psi)$, but I will assume[32]

$$\mathsf{pre}(\varphi_1, \ldots, \varphi_n \Rightarrow \psi) := \Box(\varphi_1 \wedge \cdots \wedge \varphi_n \rightarrow \psi),$$

so not only can an agent deduce conclusions that follow logically from premises, as in

$$\begin{array}{ll}
\varphi \wedge \psi \Rightarrow \varphi & \text{and} \\
\varphi, \varphi \rightarrow \psi \Rightarrow \psi, &
\end{array}$$

for which

$$\begin{array}{ll}
\Box((\varphi \wedge \psi) \rightarrow \varphi) & \text{and} \\
\Box((\varphi \wedge (\varphi \rightarrow \psi)) \rightarrow \psi) &
\end{array}$$

are valid,[33] but the agent can even "deduce" conclusions that are strictly implied by (but do not follow logically from) the premises, as in

$$\texttt{red} \Rightarrow \texttt{colored},$$

where `red` and `colored` are atomic sentences such that

$$\Box(\texttt{red} \rightarrow \texttt{colored})$$

[32]Here one may wish to impose a constraint on our models mentioned after Definition 5.1, that if $v \in W_w$, then $W_v = W_w$, which implies that the precondition is met at w just in case it is also met at all worlds possible relative to w. However, this will not matter for our purposes.

[33]Therefore, $K\Box((\varphi \wedge \psi) \rightarrow \varphi)$ and $K\Box((\varphi \wedge (\varphi \rightarrow \psi)) \rightarrow \psi)$ are valid, assuming the **enough** condition, so the agent knows the entailments. I have not, however, built into the precondition of the deduction that the agent must know the entailment to perform the deduction.

is true at our given pointed model.[34]

With this setup, we can represent different views about the extent to which competent deduction is *guaranteed* to extend knowledge. For example, recall that over MA models satisfying **cover** and $\mathsf{RO}_{\exists\forall}$, the rule

$$\text{RM } \frac{\varphi \to \psi}{K\varphi \to K\psi}$$

is sound. Over MDA models satisfying only **cover**, this is not the case. However:

Proposition 5.8 (Single-Premise Deductive Logical Closure)**.** The following dynamic analogue of RM is sound over MDA models satisfying **cover**:

$$\text{DRM } \frac{\varphi \to \psi}{K\varphi \to \langle \varphi \Rightarrow \psi \rangle K\psi}.$$

Proof. We first observe that for any model $\mathfrak{M}$ and formulas ψ and φ,

$$[\![\psi]\!]^{\mathfrak{M}} = [\![\psi]\!]^{\mathfrak{M}_{\varphi\Rightarrow\psi}}, \tag{5.8}$$

which means that deducing ψ does not change the truth value of ψ. The argument is a simple induction on the structure of ψ. Suppose ψ is of the form $K\chi$. To show: for all $w \in W$, $\mathfrak{M}, w \vDash K\chi$ iff $\mathfrak{M}_{\varphi\Rightarrow\psi}, w \vDash K\chi$. By Definition 5.10, since $\chi \neq \psi$, we have

$$\mathrm{u}_{\varphi\Rightarrow\psi}(\chi, w) = \mathrm{u}(\chi, w). \tag{5.9}$$

By the inductive hypothesis, $[\![\chi]\!]^{\mathfrak{M}} = [\![\chi]\!]^{\mathfrak{M}_{\varphi\Rightarrow\psi}}$, which with (5.9) implies that there is some $A \in \mathbf{r}([\![\chi]\!]^{\mathfrak{M}}, w)$ such that $A \cap \mathrm{u}(\chi, w) = \emptyset$ iff there is some $A \in \mathbf{r}([\![\chi]\!]^{\mathfrak{M}_{\varphi\Rightarrow\psi}}, w)$ such that $A \cap \mathrm{u}_{\varphi\Rightarrow\psi}(\chi, w) = \emptyset$, which gives $\mathfrak{M}, w \vDash K\chi$ iff $\mathfrak{M}_{\varphi\Rightarrow\psi}, w \vDash K\chi$.

Now to the main part of the proof, illustrated in Fig. 5.14 below: if $\varphi \to \psi$ is valid, then for any model $\mathfrak{M}$,

$$[\![\varphi]\!]^{\mathfrak{M}} \subseteq [\![\psi]\!]^{\mathfrak{M}}, \tag{5.10}$$

[34] As observed in footnote 15, if we assume both **r-possible** and **enough**, then $\Box\alpha \to K\alpha$ is true, so the agent will know the implication `red` $\to$ `colored`. I have not, however, built into the precondition of the deduction that the agent must know the implication to perform the deduction.

so by **cover** we have

$$\forall B \in \mathbf{r}([\![\varphi]\!]^{\mathfrak{M}}, w)\ \exists A \in \mathbf{r}([\![\psi]\!]^{\mathfrak{M}}, w)\colon A \subseteq B. \tag{5.11}$$

Now suppose $\mathfrak{M}, w \vDash K\varphi$, so there is some $B \in \mathbf{r}([\![\varphi]\!]^{\mathfrak{M}}, w)$ such that

$$B \cap \mathrm{u}(\varphi, w) = \emptyset. \tag{5.12}$$

It follows by (5.11) that there is some $A \in \mathbf{r}([\![\psi]\!]^{\mathfrak{M}}, w)$ such that

$$A \cap \mathrm{u}(\varphi, w) = \emptyset. \tag{5.13}$$

By Definition 5.10 and the assumption that $\mathfrak{M}, w \vDash K\varphi$,

$$\mathrm{u}_{\varphi \Rightarrow \psi}(\psi, w) = \mathrm{u}(\psi, w) \cap \mathrm{u}(\varphi, w). \tag{5.14}$$

It follows by (5.13) and (5.14) that

$$A \cap \mathrm{u}_{\varphi \Rightarrow \psi}(\psi, w) = \emptyset. \tag{5.15}$$

Given (5.8), $A \in \mathbf{r}([\![\psi]\!]^{\mathfrak{M}}, w)$ implies $A \in \mathbf{r}([\![\psi]\!]^{\mathfrak{M}_{\varphi \Rightarrow \psi}}, w)$, with which (5.15) implies $\mathfrak{M}_{\varphi \Rightarrow \psi}, w \vDash K\psi$. Finally, since $\varphi \to \psi$ is valid, $\Box(\varphi \to \psi)$ is also valid, so the precondition in Definition 5.11 is satisfied and we have $\mathfrak{M}, w \vDash \langle \varphi \Rightarrow \psi \rangle K\psi$. □

Proposition 5.8 shows that in my proposed version of the Multipath Picture of Knowledge with the Five Postulates of §5.2.1, if an agent knows φ—relative to an attributor's context $\mathcal{C}$—and competently deduces a logical consequence ψ from φ, then the agent is *guaranteed* to know ψ—relative to the same context $\mathcal{C}$. Contextualists may claim that in certain cases, the attributor's attending to the fact that the agent has deduced ψ will change the attributor's context, but that is a separate issue. If an agent makes a deduction, unbenownst to the attributor, then her doing so does not automatically change the attributor's context. We must be careful not to confuse the dynamics of deduction, which happens on the side of the knowing agent, with the

dynamics of context change, which happens on the side of the attributor.

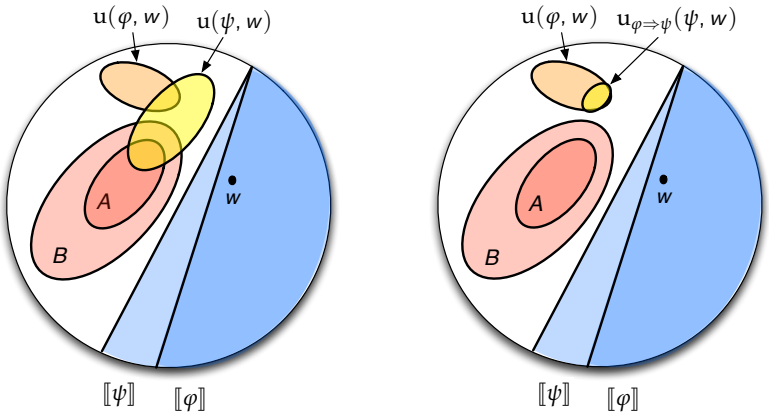

Figure 5.14: illustration for the proof of Proposition 5.8

We can prove analogues of Proposition 5.8 for other closure principles. For example, recall that over MA models satisfying **cover**, **combine**, and $\mathsf{RO}_{\exists\forall}$, the axiom

$$\text{K} \quad (K\varphi \wedge K(\varphi \to \psi)) \to K\psi$$

is valid. Over MDA models satisfying **cover** and **combine**, this is not the case. However:

Proposition 5.9 (Deductive Closure Under Known Implication)**.** The following dynamic analogue of K is sound over MDA models satisfying **cover** and **combine**:

$$\text{DK} \quad (K\varphi \wedge K(\varphi \to \psi)) \to \langle \varphi, \varphi \to \psi \Rightarrow \psi \rangle K\psi.$$

The proof is similar to that of Proposition 5.8 and left to the reader.

One can multiply results like Propositions 5.8 and 5.9, but the message is clear: for each *static* closure principle, we can consider the corresponding *dynamic*, *deductive* closure principle; and just as the validity of a static closure principle over MA models depends on properties of the **r** function, the same applies to the validity of the corresponding dynamic, deductive closure principle over MDA models. I leave discussion of further technical aspects of this dynamic deductive logic to future work.

It is also important to observe that static, *pure* closure principles (recall Remark 2.1) still make sense in this framework. For example, in connection with the view of elimination suggested above, it seems that any inquiry that concerns the question of φ *and* ψ concerns the question of φ, so if a $\neg\varphi$-world gets eliminated as an alternative to $\varphi\wedge\psi$, it is also gets eliminated as an alternative for φ. In other words, we should adopt the constraint that $\mathrm{u}(\varphi, w) \subseteq \mathrm{u}(\varphi \wedge \psi, w)$, which together with **cover** guarantees the validity of the pure closure principle $K(\varphi \wedge \psi) \to K\varphi$, and similarly for ψ. Moreover, it seems that any inquiry that concerns the question of $\varphi \wedge \psi$ concerns the question of $\psi \wedge \varphi$, and so on (see the discussion of "syntactic variants" in §6.2.2). Constraints on the u function (together with **cover**) give rise to pure closure principles that can live alongside dynamic deductive closure principles in the Transfer Picture.

At last, we have a model of the epistemic effect of "putting two and two together" that is so important for us non-ideally astute logicians. Whether putting two and two together in a certain way always guarantees the extension of knowledge depends on one's theory of the **r** function. But it is important to note that fallibilists with different theories of the **r** function can still accept the Transfer Picture of Deduction, given by Definition 5.10, just as they can still accept the Multipath Picture of Knowledge.

According to my theory of the **r** function, given by the Five Postulates in §5.2.1, the principle DK of deductive closure under known implication is not valid. This is because I do not assume that the **combine** postulate holds for *every* pair of propositions. For the same reason, the principle $(K\varphi \wedge K\psi) \to \langle \varphi, \psi \Rightarrow \varphi \wedge \psi\rangle K(\varphi \wedge \psi)$ is not valid. Yet competently deducing $\varphi \wedge \psi$ from known premises φ and ψ *sometimes*—indeed *often* for those who stay away from tricky propositions—results in knowledge of $\varphi \wedge \psi$, because the **r** function satisfies the instance of **combine** for the particular propositions $[\![\varphi]\!]^{\mathfrak{M}}$ and $[\![\psi]\!]^{\mathfrak{M}}$. Let us not forget that *validity* is a strong thing. In Chapter 6, I will continue to defend the view that when it comes to these multi-premise closure principles, validity is something we fallibilists can live without.

5.5 Conclusion

In this chapter, I have presented new fallibilist views of knowledge and deduction, the Multipath Picture of Knowledge and the Transfer Picture of Deduction. My motivating arguments for these pictures in §5.1.1, §5.1.2, and §5.4 were aimed at all fallibilists, whatever their views on disputed issues like closure and contextualism. What these pictures revealed in §5.2.3 and §5.4 is that the problems of Vacuous Knowledge, Containment, and Knowledge Inflation from Chapter 4 are artifacts of the standard but flawed framework of Fallibilism 1.0, not problems inherent in fallibilism itself. Finally, in §5.3 and §5.4, I staked a position on the disputed issue of deductive closure for fallibilists: single-premise logical closure fits well with fallibilism, while multi-premise logic closure and closure under non-logical (bi-)entailment push fallibilists in the wrong direction, back to vacuous knowledge. In Chapter 6, I will further defend my position on closure—from both those who think I have not admitted enough closure, as well as those who think I have admitted too much.

Of course, there is much more to be done to realize the blueprint of the Multipath Picture of Knowledge in a specific fallibilist theory of knowledge. Like other fallibilists, I have appealed to the idea that some possibilities need not be eliminated in order to know some proposition P relative to a context $\mathcal{C}$, without giving a criterion for a possibility to be "irrelevant" to knowing P relative to $\mathcal{C}$ in that sense. As I suggested at the end of §5.3 and I will show in other work,[35] there are various ways to generate a multipath function $\mathbf{r}$ from a family of world-orderings thought of in any of the traditional ways, or from a set-selection function r as in Chapter 3; so the fallibilist who adopts the Multipath Picture is no worse off than traditional fallibilists with respect to explaining what makes some possibilities "irrelevant" to knowing P. Nonetheless, some kind of explanation needs to be given, and similarly for the notion of elimination, the interpretation of which is constrained by the Transfer Picture of Deduction as discussed in §5.4. I am confident that there are adequate explanations, but I will not give them here. Instead, I refer to Lawlor [2013] for discussion of "reasonable alternatives" and to Pryor [2001, §1.2] for discussion of "ruling out."

[35] Note added in ILLC version: see Holliday 2013b.

6

Objections and Replies

Here is a recap of our story so far: in Chapter 5, the Multipath Picture of Knowledge and Transfer Picture of Deduction resolved the Problems of Vacuous Knowledge, Containment, and Knowledge Inflation raised in Chapter 4 for the framework of Fallibilism 1.0 proposed in Chapter 3 as a generalization of the RA and subjunctivist theories from Chapter 2. As argued in §5.3, the complete resolution of the Problem of Vacuous Knowledge for fallibilism required rejecting the validity of closure under known implication—and hence multi-premise closure—as well as closure under known strict (bi-)implication, the key principles used in closure-based arguments for radical skepticism about knowledge. However, as shown by the consistency of the Five Posulate in §5.2.3 and the Transfer Picture in §5.4, fallibilists can maintain single-premise closure without either vacuous knowledge or knowledge inflation.

In this chapter, I consider objections to my position on closure that accepts single-premise logical closure but rejects the stronger principle(s) of closure under known (strict) (bi-)implication. In §6.1.1 and §6.1.2, I consider objections according to which I have not guaranteed *enough* closure, while in §6.2, I consider objections according to which I have guaranteed *too much* closure. In each case, I argue that the objection fails. The reasons for their failure illuminate the relation between closure and the pragmatics of assertion, the structure of skeptical arguments, and the relation between closure and the logical strength and subject matter of what we know.

6.1 Not Enough Closure

In this section, I consider two arguments according to which my position on closure, which rejects closure under known (strict) (bi-)implication, does not guarantee enough closure. I will assume from the beginning that rejecting full closure is preferable to accepting skepticism. Luper-Foy [1987] remarks that "if the rejection of skepticism depends on the rejection of the Entailment Principle, then perhaps we would be better off adopting skepticism" (6). How so? Adopting skepticism means adopting the view that we know almost nothing about the world, that our ordinary practices of ascribing knowledge are radically in error. It turns our view of the relation between ourselves and the world upside-down. The violence done to our view of ourselves and the world by rejecting closure under known implication, a rarified philosophical principle, pales in comparison. The idea that we would be better off, in the face of the "skeptical paradox" [Cohen, 1988], to accept that we know nothing rather than give up unrestricted closure under known implication strikes me as implausible as the idea that we would be better off, in the face of Russell's paradox, to accept true contradictions rather than give up unrestricted set comprehension.[1]

I assume that the mentality of "Give me Closure, or Give me (Epistemic) Death!" does not depend on a confusion between, on the one hand, the denial that agents *always* know what they know to be implied by (or what they "competently deduce" from) what they know, and on the other hand, the claim that agents *never* know those things. However, consider the following passage from BonJour 1987:

> [If] knowing something does not allow one to reason from it via a known entailment to some further conclusion and thereby know the result, then what, one might well ask, is the point of having knowledge in the first place? In particular, if we infer from our knowledge that a particular course of action is the best choice in a particular situation, we will not

[1] In a way, the first idea may be even more radical, since the epistemologist who wishes to retain closure at the cost of skepticism must be willing to reject almost all of our thought about knowledge, whereas the set theorist who wishes to retain comprehension at the cost of true contradictions will argue that by adopting a paraconsistent logic, we may contain the few true contradictions without too much damage to mathematical thought [Priest, 2011].

> thereby know that it is best, leaving it quite unclear how knowledge can serve the crucial role of guiding action. (310)

Of course, the phrases 'does not allow' and 'we will not' mischaracterize the position of fallibilists why deny full closure. The phrase 'does not allow' should be something like 'does not *guarantee* that one can' and 'we will not' should be something like 'we are not guaranteed, in virtue of a general closure principle, to . . . '. BonJour mentions in an endnote the crucial fact that Nozick does not claim that we *never* know things in virtue of their following from things we know via known implications,[2] and that Nozick even discusses "which formal rules always preserve knowledge" (313n22) (although I would not say that a restricted closure principle states that a "formal rule preserves knowledge," which is a confusing expression). However, BonJour claims "this does not seem to help very much . . . since very many actual entailments fail to fall neatly under formal rules" (ibid.). But fallibilists who deny full closure do not claim that in cases where none of their restricted closure principles applies, the agent fails to know the conclusion of her reasoning. It is rather that she may well know the conclusion, but this was not guaranteed in virtue of some general closure principle. Obviously it does not follow from this that there is no point to having knowledge.

6.1.1 Hawthorne on Assertion

Let us call a fallibilist who denies closure under known implication against the skeptic a "robust fallibilist" [Pritchard, 2005, 35]. The first objection I will consider concerns the *assertions* that robust fallibilists are supposedly willing to make. The objection has two parts, the "abominable conjunction" objection [DeRose, 1995] and what we might call the "Tortoise" objection [Hawthorne, 2004a]. We begin with the second, as Hawthorne explains it:

> [D]enial of closure interacts disastrously with the thesis that knowledge is the norm of assertion. One who attempts to conform to this norm but simultaneously attempts to adhere to the Dretske-Nozick strategy will end

[2] "[W]e have not said that . . . knowledge never flows down from known premises to the conclusion known to be implied, merely that knowledge . . . does not always flow down" [Nozick, 1981, 230].

> up behaving rather like a familiar object of ridicule—Lewis Caroll's (1895) Tortoise. For if you place such a character in front of a zoo cage containing a zebra and ask him 'Do you agree that the thing in the cage is a zebra?', he will say 'Yes', and if you ask him, 'Do you also agree that if the thing in the cage is a zebra, then the thing in the cage is not a cleverly disguised mule?', he will also say 'Yes'. But if you then ask him, 'So you agree that the thing in the cage is not a cleverly disguised mule?', he will now say 'Oh no. I'm not agreeing to that'. (By his lights, he does not know the conclusion and thus, given the controlling norm, will not assert it.) Now sometimes when we have a consequence of our beliefs pointed out to us, we do not embrace the consequence. Rather, given the unpalatability of the consequence, we give up one of the original beliefs. But that is not what is going on here either. The premises of the modus ponens argument are stably adhered to and yet the conclusion is stably repudiated. (39)

Although there is controversy (see Weiner 2007) surrounding the idea that knowledge is the norm of assertion, that one must assert a proposition only if one knows the proposition [Williamson, 2001, Ch. 11], let us grant it for the sake of argument. Even so, it does not follow that the robust fallibilist would "repudiate" the conclusion of the modus ponens argument, responding "I'm not agreeing to that," as in Hawthorne's caricature. It does not follow for several reasons. Harman and Sherman [2004] clearly explain one of the mistakes involved in this form of the Tortoise objection:

> Hawthorne argues that this can lead to an odd conversation: Alice asserts that the animal in the cage is a zebra and agrees that, if the animal is a zebra, then it is not a cleverly disguised mule; however, she is not willing to agree that the animal is not a cleverly disguised mule. But this is a mistake. Alice *accepts* that the animal is not a cleverly disguised mule. In fact, she assumes that. She just doesn't take herself to know it and so does not assert it, although she can assert that it is something she accepts. So, we see no difficulty here. (496-497)

Hawthorne [2004b, 512] still feels uncomfortable that Alice would be unwilling to just answer "Yes" when asked if the conclusion is true. But it would be natural for her to reply "I think so," which hardly makes her an object of ridicule like Caroll's Tortoise. On the other hand, what does it mean for Alice to "stably adhere" to the premise that the thing in the cage is a zebra? Does it mean that she must be willing to re-assert that it is a zebra, after someone has just raised to her the skeptical hypothesis of the cleverly disguised mule? Even those who do not deny closure recognize pragmatic reasons for why making the same type of utterance would be inappropriate after one's interlocutor raises a skeptical hypothesis (e.g., Pritchard 2005, §3.2; Turri 2010), and robust fallibilists can recognize them too. But before turning to these pragmatic issues, let us consider an objection related to the Tortoise objection.

Suppose that in the passage from Hawthorne, we replace the phrase 'Do you agree that ...' by the phrase 'Do you agree that *you know* ...'. Hence we have moved up one level epistemically. Now the principle behind Hawthorne's questions is not *modus ponens*, but *closure under known implication*. The objection becomes: if Alice is a robust fallibilist, will she not end up "stably adhering" to the premise that she knows the thing is a zebra while "stably repudiating" the conclusion that she knows the thing is a cleverly disguised mule. Is this not "abominable" [DeRose, 1995]?

I ask again: what does it mean for her to "stably adhere" to the premise? Must she be willing to re-assert that she knows the thing in the cage is a zebra, after someone has just raised to her the skeptical hypothesis of the cleverly disguised mule? These questions demand that we consider the pragmatics of dialogue with a skeptic.

Two Kinds of Skeptical Dialogue

For reasons explained in §6.2, I will now shift from Dretske's zebra vs. painted mule case to his Gadwall vs. Siberian Grebe case. Suppose that a fallibilist park ranger and a skeptical ornothologist are talking about Alice, who is standing with her son in front of the lake with the Gadwall on it. Further suppose that the ranger and ornothologist have just inspected the Gadwall, so there is no question between them about what kind of bird it is. They can say definitively that it is a Gadwall, not a Siberian Grebe, not an animatronic robot, etc. The following dialogue ensues:

DIALOGUE I

Ornothologist: Does she know it's a Gadwall?

Ranger: Yes, I overheard her talking to her son. She knows it's a Gadwall. In fact, she even identified it as a female Gadwall.

Ornothologist: Wow, she got that right. But does she really know it's a Gadwall?

Ranger: What do you mean?

Skeptic: Well, does she know it isn't a Siberian Grebe or an animatronic robot planted to fool tourists or . . . [he lists other skeptical hypotheses]?

Ranger: Oh come on, that's not necessary for her to know it's a Gadwall, which she does.

For comparison with DIALOGUE I, consider a dialogue between Alice and her son:

DIALOGUE II

Son: What kind of bird is that?

Alice: It's a Gadwall. In fact, you can tell from the plumage that it's a female Gadwall.

Son: Do you really know that?

Alice: What do you mean?

Son: Well, do you know it isn't a Siberian Grebe or an animatronic robot planted to fool tourists or . . . [he lists other skeptical hypotheses]?

Alice: . . .

Some might expect Alice, if she is a robust fallibilist, to reply to her son as the ranger did to the ornothologist, with "Oh come on, that's not necessary for me to know it's a Gadwall, which I do" or with "Oh come on, that's not necessary for me to know that it's a Gadwall, which it is," and leave it at that. However, there is an important difference between the two dialogues for the robust fallibilist.

The role of the son's question about the possibility of the Siberian Grebe in DIALOGUE II is not the same as the role of the ornothologist's question about it in DIALOGUE I. The ornothologist asks the question in order to raise doubt about whether Alice knows that the bird is a Gadwall, but not to raise doubt about whether the bird *is* a Gadwall. There is no question in DIALOGUE I that Alice's belief about the Gadwall is true; the ornothologist only questions whether her true belief constitutes knowledge. However, at the beginning of DIALOGUE II, it is not part of the common ground (in the sense of Stalnaker 2002) that the bird is a Gadwall, as shown by the son's first question. Hence when the son asks about the Siberian Grebe possibility, he raises doubt about whether the bird is in fact a Gadwall. For this reason, it would be inappropriate for Alice to reply to her son in the way that the ranger replies to the ornothologist. For her son is in effect questioning something that *is* a necessary condition for her Gadwall knowledge: that the bird isn't a Siberian Grebe. By contrast, the ornothologist, who knows that this necessary condition for Alice's Gadwall knowledge is satisfied, is not questioning it. He is questioning something else, which is *not*—according to the robust fallibilist—a necessary condition for Alice's Gadwall knowledge: that she knows that the bird isn't a Siberian Grebe. This is why the ranger's reply to the ornothologist is appropriate by fallibilist standards.

Abominable Conjunctions

The above distinctions help to explain why robust fallibilists would not assert "abominable conjunctions" [DeRose, 1995] of the form: I know it's a Gadwall, but I don't know it's not a Siberian Grebe. Robust fallibilists can recognize pragmatic reasons for why these and related assertions are inappropriate, even if true.[3] For one thing,

[3]Heller [1999a], Dretske [2005], and Sherman and Harman [2011] offer explanations for why assertions of abominable conjunctions are inappropriate, consistent with their sometimes being true.

the assertion of the second conjunct suggests that there are now special reasons for considering relevant a possibility normally assumed to be irrelevant by someone who asserts the first conjunct (cf. Dretske 2005). Why else bring it up? To this we add that an assertion of the second conjunct normally raises doubt about whether the bird is *in fact* a Gadwall, which is a necessary condition for the truth of what is asserted with the first conjunct. In Gricean [1989] terms, the first tension arises given the assumption that the speaker is observing the maxim of Relation ("Be relevant"), while the second tension arises given the assumption that the speaker is observing the supermaxim of Quality ("Try to make you contribution one that is true").

According to the pragmatic explanation, if we explicitly block the problematic suggestions, then the resulting assertion should sound better. Suppose that the ranger tells the ornothologist about Alice, "Yes, she knows that our Gadwall is a Gadwall—in fact, she even knows it's a female Gadwall—although she of course hasn't ruled out [or doesn't know to be false] your wild fantasies about it's being a Siberian Grebe or an animatronic robot."[4] Is the ranger's assertion debatable? Yes. Abominable? No. If we are honest, Lewis [1996, 549f] says, then we will admit that saying "He knows, yet he has not eliminated all possibilities of error" sounds wrong, even contradictory. But as Lewis [1973] writes in a different context, "oddity is not falsity; not everything true is a good thing to say. In fact, the oddity dazzles us. It blinds us to the truth value of the sentences" (28). Try this instead: "he knows, for what he believes is true and he has eliminated all possibilities of error—except, of course, the *bizarre* ones" (cf. Rysiew 2001, 495). As before, this is debatable, but not abominable.[5]

To block the second problematic suggestion of the "abominable conjunction" (the doubt that the bird is in fact a Gadwall), we switched to assertions in the third

Unlike the explanations given by Dretske and Sherman and Harman, Heller's explanation appeals to contextualism. There is a related literature, started by Rysiew [2001] and Stanley [2005], on *concessive knowledge attributions* (CKAs) such as "I knows it's a Gadwall, but it's possible that it's a Siberian Grebe." But neither Rysiew nor Stanley attempts to apply his explanation of the infelicity of CKAs to abominable conjunctions. In fact, Stanley [2005] thinks that the infelicity of asserting the CKA above is due to the fact that this CKA is always *false* (for a different kind of CKA, Stanley offers a pragmatic explanation instead), but he also argues that fallibilism is not committed to its truth; he has the same view on abominable conjunctions, since he accepts epistemic closure.

[4]The intended reading is, of course, that Alice knows that *our Gadwall* (de re) *is a Gadwall*.

[5]Also see Pritchard 2005, 89 on the pragmatic awkwardness of uttering Lewis's sentence.

person for a reason. Since the ranger and ornothologist have settled that the bird is a Gadwall, not a Siberian Grebe, the ranger can make his assertion against this background. However, the issue is not settled in the same way in a conversation between you and me, if I admit to you that I do not know that the bird is not a Siberian Grebe. The problem is not just the use of the first person, but the use of the present tense. For I can admit that I *did not* know that it was not a Siberian Grebe, while maintaining in robust fallibilist fashion that I knew it was a Gadwall. For example, suppose that after talking to the ranger, the ornthologist approaches Alice, and the following dialogue ensues.

DIALOGUE III

Ornothologist: Do you know what kind of bird that is?

Alice: It's a Gadwall. In fact, you can tell from the plumage that it's a female Gadwall.

Ornothologist: That's right. But do you really know it's a Gadwall?

Alice: Yes, you just told me it is.

Ornothologist: What I meant to ask was—did you really know it was a Gadwall, before I told you?

Alice: What do you mean?

Ornothologist: Well, did you know it wasn't a Siberian Grebe or an animatronic robot planted to fool tourists or . . . [he lists other skeptical hypotheses]?

Alice: Oh come on, that wasn't necessary for me to know it was a Gadwall, which I did.

By the middle of this dialogue, it is common ground that the bird is indeed a Gadwall. Hence the role of the ornothologist's question about the possibility of the Siberian Grebe is the same as in DIALOGUE I, as explained above. This is why the robust fallibilist takes Alice's response in DIALOGUE III to be appropriate.

Application to Global Skepticism

Let us now apply what we have observed to the case of *global* skepticism. The important difference from the Gadwall case is that it is typically not common ground between the global skeptic and the non-skeptic that (for example) there is definitely a mind-independent world, the only question being whether some person knows this fact. Perhaps skepticism is treated in this way in some seminar rooms. However, there is a kind of philosophical skeptic who questions whether anyone—himself included—knows that there is an external world in part by questioning whether *there is* one, much like the son in DIALOGUE II in effect questions whether there is a Gadwall on the lake. Receiving no answer that satisfies him, this kind of skeptic concludes that we do not know the ordinary things about the world that we purport to know.

One might have thought that if we can deny closure, then we can insist against the challenge of the global skeptic that we *do* know the ordinary propositions in question.[6] However, our analysis of DIALOGUE II suggests that insisting on this (even if true) is not an appropriate answer to the skeptic's line of questioning.[7] We should be more modest. What the denial of closure allows us to say to the global skeptic is that his conclusion that we know next to nothing does not follow from his premise that we do not know that all of his global skeptical hypotheses are false. It is consistent with his argument that we know a great deal about the world (not

[6]Referring to a scenario in which a skeptic raises the possibility that what appear to be oranges are instead wax imitations, Dretske [2004, 40] writes: "Agreeing with the skeptic ... that I don't now, and never did, know they aren't wax leaves me (unlike a radical contextualist) free to insist that I nonetheless knew what I then said I knew—that they were oranges. .... *That*, it seems to me, is a meaningful answer to skepticism." Dretske apparently has in mind a situation not like that of DIALOGUE III, where it has become common ground that what appeared to be oranges were in fact oranges, but rather like that of DIALOGUE II, where it has not (given the "I don't now ..."). I consider the more modest reply to the skeptic given in the main text to be a meaningful answer to skepticism. Also see note 8.

[7]This does not mean that asserting knowledge of ordinary propositions is inappropriate in other contexts. That the same knowledge-ascribing utterance can be an appropriate speech act in one context and an inappropriate one in another context can be explained in several ways. The standard explanation is that what counts as an appropriate assertion is context-dependent. An alternative explanation, proposed by Turri [2010], is that by uttering the same knowledge-ascribing sentence in different contexts, one may perform *different kinds* of speech acts, e.g., an *assertion* in an ordinary context and a *guarantee* in a skeptical context, and these different kinds of speech acts have different standards of appropriateness.

only possibly, but actually). That being said, our denial that knowledge has a certain closure property should hardly reassure someone worried about whether the world has existed for more than five minutes, or whether there is a mind-independent world at all, which are necessary conditions for the *truth*—and hence our knowledge—of many ordinary propositions. This is why it seems inappropriate to baldly claim knowledge of such propositions against the skeptic (even if we have it) and leave it at that. We are left in an uncomfortable position. If anyone thought there was an easy way out, they have failed to appreciate the force of philosophical skepticism.

One should not mistake the call for modesty against the skeptic for a concession that we do not *know that we know* the ordinary propositions disputed by the skeptic.[8] If we are externalists about knowledge, as many robust fallibilists are, then we should allow that we may know without knowing that we know, that we may know that we know without knowing that we know that we know, etc. (see Dretske 2004), and even that we do not know up to which level we know (see Nozick 1981, 247). However, as robust fallibilists, we also maintain that nothing the skeptic has said establishes that we do not have a great deal of higher-order knowledge. Arguments to the effect that we do not depend on closure as much as arguments against first-order knowledge do.

How does the robust fallibilist compare in dialogue with a skeptic to a neo-Moorean (recall Remark 2.5) or contextualist? According to Pritchard [2005, §3.2-3.4], neo-Mooreans should not emulate Moore's boldness against the skeptic. For Moore's claims of knowledge—even of ordinary propositions—in the face of the skeptic were (true but) pragmatically inappropriate. Much of Pritchard's explanation of this

[8]Those who hold that knowledge is *sufficient* for appropriate assertion may interpret the call for modesty in this way. I join Pritchard [2005, §3.2-3.4] in thinking that the sufficiency thesis fails for skeptical conversations. (For other arguments against sufficiency, see Brown 2010.) In taking this position, I depart from some robust fallibilists. For example, McGinn [1984, 28] writes: "Non-closure allows us to accept that the sceptic's initial contention has force without being committed to the alarming conclusion that our ordinary knowledge claims are false. This seems to me some advance against the sceptic, but it leaves an important question open: do we *know* that our ordinary knowledge claims are true? We commonly think, not only that we may have knowledge that there is a table there, but also that we know that we do—so that we are in a position to *assert* that we know that there is a table there. It therefore seems that if we are to be at all consoled by the antisceptical consequences of non-closure, we need to sustain our conviction that we know that we know."

context-sensitive inappropriateness could be adopted by robust fallibilists and applied to the Tortoise and abominable conjunction objections as well.[9]

As for contextualists, they can be no bolder against the skeptic. For their "solution" to skepticism (see, e.g., Cohen 1988, DeRose 1995, Lewis 1996) does not allow them to insist, when challenged in dialogue with a radical skeptic, that they know ordinary propositions either—or even that they know such propositions according to *ordinary standards* of knowledge. This is a consequence of the much-discussed Problem(s) of Factivity for contextualism [Williamson, 2001, Wright, 2005, Brendel, 2005, Baumann, 2008] (see Appendix §6.A). In fact, this problem shows that contextualists must concede even more in dialogue with a skeptic, namely that they *do not* know that they count as knowing ordinary propositions relative to ordinary standards of knowledge.[10] By contrast, as noted above, robust fallibilists make no such concession.

What the analysis of this section shows is that the objection from assertion underestimates the sensitivity of robust fallibilists to the pragmatics of dialogue with a skeptic. Anyone pushed into a corner by a skeptic ends up in an uncomfortable place. Yet the robust fallibilist, who offers a serious critique of the skeptical argument, does not end up behaving in conversation in a ridiculous or abominable manner.

[9]However, I do not follow Pritchard in thinking that if Alice claims to know the negations of all of the skeptic's hypotheses, then she is *merely* saying something conversationally inappropriate (but true). In my view, if she has not done the epistemic work required to rule out skeptical possibilities, then she does not know the negations of all of the skeptics hypotheses, contrary to Pritchard's safety-based view that allows vacuous knowledge of these propositions (recall §4.1).

[10]At least they must concede this if the skeptic induces a context $\mathcal{S}$ like that described in Appendix §6.A. The contextualist may reply that the skeptic does not always succeed in inducing such a context. Hence one contextualist suggests that "the prospect of hard-nosed sceptics turning one's context into a defective context should not bother the anti-sceptic too much: sceptics are easily excluded from one's conversation by simply ignoring them" [Blome-Tillmann, 2009, 274]. I leave it to the reader to judge which is the more philosophically satisfying response in a dialogue with a skeptic, the robust fallibilist response—*argue that the skeptic's reasoning is invalid*—or this contextualist response—*wear earplugs*.

6.1.2 Hawthorne on Equivalence

In this section, I consider another argument to the effect that I have not guaranteed enough closure. Recall from §5.3 that I have rejected the closure principle

$$\text{EP} \quad (K\varphi \wedge K\Box(\varphi \leftrightarrow \psi)) \rightarrow K\psi,$$

since it leads to the Problem of Vacuous Knowledge or the Problem of Containment. Hence I reject the closure step in (5) of the following skeptical argument discussed by Hawthorne [2004a, 41]. Assume s is a *strict* counter-hypothesis to p (Remark §5.3), so $\Box(p \rightarrow \neg s)$ holds, and to reduce symbols let $\varphi \leftrightarrows \psi$ stand for $\Box(\varphi \leftrightarrow \psi)$. Each line contains the following from left to right: line number, formula, justification, set of open assumptions, and my evaluation ($\times$ for rejection and ✓ for endorsement).

(1)	$\neg K\neg s$	premise	{(1)}	granted
(2)	$K(p \wedge \neg s) \rightarrow K\neg s$	M	{}	✓
(3)	$\neg K(p \wedge \neg s)$	(1), (2), PL	{(1)}	granted
(4)	$K(p \leftrightarrows (p \wedge \neg s))$	premise	{(4)}	granted
(5)	$(Kp \wedge K(p \leftrightarrows (p \wedge \neg s))) \rightarrow K(p \wedge \neg s)$	EP	{}	$\times$
(6)	$\neg Kp$	(3) - (5), PL	{(1), (4)}	$\times$

I also reject (9) in the following skeptical argument from Hawthorne [2004a, 41], where AC is the principle $K\varphi \rightarrow K(\varphi \vee \psi)$, which Hawthorne calls "addition closure":

(7) $\neg K\neg s$	premise	$\{(7)\}$	granted
(8) $K((p \vee \neg s) \,\text{ɛ⊐}\, \neg s)$	premise	$\{(8)\}$	granted
(9) $(K(p \vee \neg s) \wedge K((p \vee \neg s) \,\text{ɛ⊐}\, \neg s)) \rightarrow K\neg s$	EP	$\{\}$	×
(10) $\neg K(p \vee \neg s)$	(7) - (9), PL	$\{(7), (8)\}$	×
(11) $Kp \rightarrow K(p \vee \neg s)$	AC	$\{\}$	✓
(12) $\neg Kp$	(10), (11), PL	$\{(7), (8)\}$	×

According to Hawthorne [2004a, 41], the principle EP is "very compelling." Given that EP leads to either the Problem of Vacuous Knowledge or Problem of Containment (§5.3), I do not find it compelling on reflection. Yet one might object that it is one thing to show that EP leads to these problems, as a basis for claiming that fallibilists should give it up, and another thing to explain *why* EP is not valid. As Hawthone observes, "the counterfactual considerations that Dretske and Nozick adduce to divorce the epistemic status of some p from its a priori consequences do not similarly divorce p from its a priori equivalents" (39-40). I agree, despite recent objections by Adams et al. [2012], for reasons explained in Appendix §6.B.

However, there are other considerations that divorce the epistemic status of p from that of some of its a priori equivalents. Let p be *b is a Gadwall* and let s be a strict skeptical counter-hypothesis like *b is a duck-hologram*. Hence p and $p \wedge \neg s$ are a priori equivalents.[11] Why is it that (on my view) knowing p does not require any looking around for a hologram projector, whereas knowing $p \wedge \neg s$ may require extra investigation of some kind when one's background information is not sufficient to rule out holograms? The answer is that $p \wedge \neg s$ goes beyond p in two closely related ways: it has a greater *logical strength* and a greater *subject matter* (I owe the latter point to Yablo, discussed below). It has a greater logical strength than p because of the conjunct with the logically unrelated $\neg s$, and it has a greater subject matter than p because $\neg s$ is *about* holograms, so $p \wedge \neg s$ is (partly) about holograms, whereas p is not at all about holograms; p is about b and a common flesh-and-blood species of

[11] I am assuming that it is knowable *a priori* that if b is a Gadwall, then b is not hologram.

duck. In short, $p \wedge \neg s$ *says more about more.* This is why it can take *more work* to know it. Indeed, this is what allows the *a priori* equivalent $p \wedge \neg s$ of p to have what Dretske [2005] calls a "heavyweight" status compared to the "lightweight" status of p.

Not only do we have a demonstration of how EP leads to either the Problem of Vacuous Knowledge or the Problem of Containment, but also we have an explanation of why EP is not valid, based on the natural fallibilist idea that the range of alternatives that one must eliminate in order to know something depends on its *logical strength* and *what it's about.* This explanation assumes that the objects of knowledge are more fine-grained than propositions understood as sets of metaphysically possible worlds, in line with my position in Chapter 5. Now I am putting this together with what Lewis [1988, §XI] calls the *hyper-intenstional*, "part-of-statements" conception of (partial) aboutness, which Lewis regards as one of a number of legitimate notions of aboutness. In my view, the idea that EP is "very compelling" is symptomatic of a dangerous general tendency. It is what Perry [1989] has called "losing track of subject matter": losing track of what propositions are *about*, considering only the possibilities in which they are true. Barwise and Perry [1983, 1996] have argued that losing track of subject matter leads to serious problems in semantics. Recently Stephen Yablo has drawn attention to the importance of keeping track of subject matter in connection with epistemic closure,[12] and the argument above for why knowing $p \wedge \neg s$ may be more difficult than knowing p is similar to Yablo's subject-matter based arguments against closure. Yet it is also different in an important respect, which leads to a divergence of my view and Yablo's. I will explain this divergence in §6.2.1.

6.2 Too Much Closure

According to the objection of the previous section, I have not admitted enough closure. According to the objection of this section, I have admitted too much closure.

[12]Yablo presented his ideas on these topics in his Kant Lectures, "Truth and Aboutness" and "Achieving Closure," at Stanford University on May 19-20, 2011.

Recall the (static) formulation of single-premise logical closure with the rule

$$\text{RM} \quad \frac{\varphi \to \psi}{K\varphi \to K\psi}.$$

The argument that single-premise logical closure is too much closure depends on the claim that the following principles, derivable from RM, are problematic:

$$Kp \to K\neg(\neg p \wedge s) \tag{6.1}$$

$$K\neg p \to K\neg(p \wedge s). \tag{6.2}$$

First observe that (6.1) and (6.2) are derivable from

$$\text{AC} \quad K\varphi \to K(\varphi \vee \psi)$$

together with closure under logical equivalence,

$$\text{RE} \quad \frac{\varphi \leftrightarrow \psi}{K\varphi \leftrightarrow K\psi},$$

as follows:

(13)	$Kp \to K(p \vee \neg s)$	AC
(14)	$(p \vee \neg s) \leftrightarrow \neg(\neg p \wedge s)$	PL
(15)	$K(p \vee \neg s) \leftrightarrow K\neg(\neg p \wedge s)$	(14), RE
(16)	$Kp \to K\neg(\neg p \wedge s)$	(13), (15), PL,

and similarly for (6.2).

The problem with (6.1) is supposed to arise when p is a mundane proposition and s is an incompatible skeptical hypothesis. For example, Nozick [1981, 229] writes:

> Also, it is possible for me to know p yet not know the denial of a conjunction, one of whose conjuncts is not-p. I can know p yet not know ... not-(not-p & SK). I know I am in Emerson Hall now, yet I do not

> know that: it is not the case that (I am in the tank on Alpha Centauri now and not in Emerson Hall).
>
> However, we have seen no reason to think knowledge does not extend across known logical equivalence.

Interestingly, only a few pages later Nozick [1981, 230] writes:

> It seems that a person can track 'Pa' without tracking 'there is an x such that Px'. But this apparent nonclosure result surely carries things too far. As would the apparent result of nonclosure under the propositional calculus rule of inferring 'p or q' from 'p', which stands to existential generalization as simplification stands to universal instantiation.[13]

What is interesting about these passages is that they are inconsistent.[14] I assume Nozick knows that $(p \vee \neg s)$ is logically equivalent to $\neg(\neg p \wedge s)$, so given his endorsement of closure under known logical equivalence, if he knew $(p \vee \neg s)$ then he would know $\neg(\neg p \wedge s)$. But he says he does not know $\neg(\neg p \wedge s)$, so he must not know $(p \vee \neg s)$. But he also says he knows p and endorses $K\varphi \rightarrow K(\varphi \vee \psi)$, so he should know $(p \vee \neg s)$.

I do not think this inconsistency was simply a mistake. Instead, I suspect that it reflects an intuition that Nozick shares with others. Let us now bring Dretske into the story. While Nozick explicitly endorses AC and explicitly rejects (6.1), Dretske explicitly endorses AC and implicitly rejects (or at least is committed to rejecting) (6.2). For the first part, Dretske [1970] says that "it seems to me fairly obvious that if someone ... knows that P is the case, he knows that P or Q is the case" (1009). For the second part, consider Dretske's [1970, 1015-1016] famous zebra example:

> You take your son to the zoo, see several zebras, and, when questioned by your son, tell him they are zebras. Do you know they are zebras? Well, most of us would have little hesitation saying that we did know this. We know what zebras look like, and, besides, this is the city zoo and the animals are in a pen clearly marked "Zebras." Yet, something's being a

[13] The second quoted sentence is from the footnote to the first sentence.

[14] Kripke [2011, 199] also discusses the inconsistency, pointed out to him by Assaf Sharon and Levi Spectre.

> zebra implies that it is not a mule and, in particular, not a mule cleverly disguised by the zoo authorities to look like a zebra. Do you know that these animals are not mules cleverly disguised by the zoo authorities to look like zebras? .... I don't think you do. In this I agree with the skeptic .... I part company with the skeptic only when he concludes from this that, therefore, you do not know that the animals in the pen are zebras. I part with him because I reject the principle he uses in reaching this conclusion—the principle that if you do not know that Q is true, when it is known that P entails Q, then you do not know that P is true.

Now I will make an assumption about Dretske's view that is not in his text. The assumption is that one can know by looking at the zebras that they are not mules:

> You take your son to the zoo, see several zebras, and, when questioned by your son, tell him they are zebras. Do you know they are not mules? Well, most of us would have little hesitation saying that we did know this. We know what mules and zebras look like, and, besides, this is the city zoo and the animals are in a pen clearly marked "Zebras."

As fallibilists, surely we should say that in ordinary cases of observing zebras at the zoo, people who know the difference between zebras and mules know that the zebras are not mules: $K\neg m$. Putting this together with Dretske's view that $\neg K\neg(m \wedge d)$, where d stands for 'the animal in the pen is cleverly disguised to look like a zebra', Dretske must deny (6.2). Moreover, since Dretske explicitly endorses $K\varphi \rightarrow K(\varphi \vee \psi)$, an instance of which is $K\neg m \rightarrow K(\neg m \vee \neg d)$, it follows that Dretske must deny $K(\neg m \vee \neg d) \rightarrow K\neg(m \wedge d)$ and hence closure under logical equivalence (RE).

Faced with fellow fallibilists like Dretske and Nozick who reject (6.1) and (6.2), we have three choices:

1. Deny AC, $K\varphi \rightarrow K(\varphi \vee \psi)$, even for IALs.

2. Deny the "De Morgan" closure principle $K(\pm\varphi \vee \pm\psi) \rightarrow K\neg(\mp\varphi \wedge \mp\psi)$[15] and hence closure under logical equivalence, RE, even for IALs.

[15]Notation: $\pm\varphi$ is either φ or $\neg\varphi$; if $\pm\varphi$ is φ, then $\mp\varphi$ is $\neg\varphi$; if $\pm\varphi$ is $\neg\varphi$, then $\mp\varphi$ is φ.

3. Defend single-premise logical closure, RM, and hence (6.1) and (6.2), at least for IALs.

I will consider these choices in 6.2.1, 6.2.2, and 6.2.3, respectively.

6.2.1 Yablo and the Denial of AC

In §6.1.2, I argued that the different epistemic status of $p \wedge \neg s$ relative to p is a result of the greater logical strength and greater subject matter of $p \wedge \neg s$, which are closely related, due to the new conjunct $\neg s$. The idea that closure should be restricted by considerations of subject matter has been proposed by Yablo, but his restrictions are stronger than mine.[16] According to Yablo, knowing a proposition lets an agent know what follows from that proposition, *barring a change in subject matter*. Hence Yablo and I agree on the claim made in §6.1.2 that closure does not necessarily get us from knowledge of p to knowledge of $p \wedge \neg s$.[17] However, Yablo also claims that closure does not necessarily get us from knowledge of p to knowledge of $p \vee \neg s$ either. According to his view, an IAL may know the logically stronger proposition without knowing the logically weaker one, because the logically weaker one brings in new subject matter with the disjunct $\neg s$. Yablo rejects AC and step (13) in the derivation of (6.1).

Yablo's [2012b] particular theory of subject matter, coupled with his theory of the relation between subject matter and epistemic closure, leads to some serious failures of single-premise logical closure. To see this, we need some definitions.

Definition 6.1 (Partial & Minimal Models)**.** Let At be a set of atomic sentences.

1. A *classical valuation* is a function $v\colon \mathsf{At} \to \{0, 1\}$.

2. A *classical model* of φ is a classical valuation v that satisfies φ ($v \vDash \varphi$), as defined in the usual way.

[16] References to Yablo without parenthetical citations are references to his Kant Lectures, "Truth and Aboutness" and "Achieving Closure," at Stanford University on May 19-20, 2011.

[17] At least it is a consequence of Yablo's formal theory of subject matter in Definition 6.2 that p and $p \wedge \neg s$ differ in subject matter. This depends on the fact that although p and $p \wedge \neg s$ are assumed to be true in the same metaphysically possible worlds, Yablo's formal theory of subject matter deals with a logical space of all classical valuations, including those that distinguish p and $p \wedge \neg s$. Whether Yablo's informal theory is supposed to allow such an inclusive logical space is not clear.

3. A *partial valuation* is a function $\mathrm{v}\colon \mathsf{S} \to \{0,1\}$ where $\mathsf{S} \subseteq \mathsf{At}$.

4. A valuation (partial or classical) $\mathrm{v}\colon \mathsf{S} \to \{0,1\}$ *extends* (resp. *strictly extends*) $\mathrm{v}'\colon \mathsf{S}' \to \{0,1\}$ iff $\mathsf{S}' \subseteq \mathsf{S}$ (resp. $\mathsf{S}' \subsetneq \mathsf{S}$) and $\mathrm{v}(p) = \mathrm{v}'(p)$ for all $p \in \mathsf{S}'$.

5. A *partial model* of φ is a partial valuation v such that for all classical valuations v extending v, $v \vDash \varphi$.

6. A *minimal model* of φ is a partial model of φ that does not strictly extend any partial model of φ.

With the help of these definitions, Yablo defines a relation $\geq$ between formulas, reading $\varphi \geq \psi$ as "φ includes ψ" or "ψ is content-part of φ."

Definition 6.2 (Yablo Inclusion)**.** Given propositional formulas φ and ψ, let $\varphi \geq \psi$ iff all of the following hold:

1. Every classical model of φ is a classical model of ψ
   (ψ is a classical consequence of φ);

2. Every minimal model of ψ is extended by a minimal model of φ
   ("φ's subject matter includes ψ's subject matter");

3. Every minimal model of $\neg\psi$ is a minimal model of $\neg\varphi$
   ("φ's subject anti-matter includes ψ's subject anti-matter").

Since Yablo defines the *overall subject matter* of a formula to be its subject matter and subject anti-matter together, conditions 2 and 3 together say that the overall subject matter of φ includes the overall subject matter of ψ. For motivation of this definition, which leads to a fascinating new theory of *content-parts*, I refer to Yablo.

According to Yablo's theory of the relation between closure and subject matter, $K\varphi \to K\psi$ (or a dynamic version thereof) is a valid closure principle only if $\varphi \geq \psi$. Observe from Definition 6.2 that $p \not\geq p \vee q$, because there is a minimal model of $p \vee q$ that is not extended by any minimal model of p: the function $\mathrm{v}\colon \{q\} \to \{0,1\}$ with $\mathrm{v}(q) = 1$. Not only that, but also $p \wedge q \not\geq p \vee q$, because there is a minimal model of $\neg(p \vee q)$ that is not a minimal model of $\neg(p \wedge q)$: the only minimal model of $\neg(p \vee q)$

is the function $\mathrm{v}'\colon \{p,q\} \to \{0,1\}$ with $\mathrm{v}'(p) = \mathrm{v}'(q) = 0$, which strictly extends a partial model of $\neg(p \wedge q)$, namely the function $\mathrm{v}''\colon \{p\} \to \{0,1\}$ with $\mathrm{v}''(p) = 0$, so v' is not a minimal model of $\neg(p \wedge q)$. What this shows is that Yablo's view implies the non-validity not only of $Kp \to K(p \vee q)$, but even of $K(p \wedge q) \to K(p \vee q)$.

In my view, the failure of such a weak closure principle as $K(p \wedge q) \to K(p \vee q)$ is an indication that something has gone wrong. One might think from this example that the problem is only with condition 3 of Definition 6.2, but condition 2 is also problematic. For it is a consequence of condition 2 that there are φ and ψ for which $K(\varphi \wedge \psi) \to K\varphi$ is *not valid* on Yablo's view. Simply observe that $(p \vee q) \wedge p \not\geq p \vee q$, because (using the same example as for $p \not\geq p \vee q$) there is a minimal model of $p \vee q$ that is not extended by any minimal model of $(p \vee q) \wedge p$: the function $\mathrm{v}\colon \{q\} \to \{0,1\}$ with $\mathrm{v}(q) = 1$. Hence $K((p \vee q) \wedge p) \to K(p \vee q)$ is not valid according to Yablo's view.[18] I take it that the failure of $K(\varphi \wedge \psi) \to K\varphi$ according to a theory of the relation between closure and subject matter is a serious strike against the theory.

I have introduced Yablo's view as a representative (in fact, the only one I know of) for the view that AC is not valid for IALs.[19] In addition, Yablo [2012a] rejects closure under (known) logical equivalence, RE, motivated in part by an example. The example is based on the reported reaction of Yablo's students when they encounter Descartes's famous Dream Argument. To set up the example, let $\texttt{MyDream}(x)$ indicate that experience x is a dream of mine, and let $\texttt{AsLifelikeAs}(x,y)$ indicate that experience x is as "lifelike" as experience y. Finally, let e refer to my total current experience. According to Yablo, if I were like many of his students, I would hold that

$$\forall x(\texttt{MyDream}(x) \to \neg\texttt{AsLifelikeAs}(x,e)), \tag{6.3}$$

but I would not claim to know just on the basis of (6.3) that

$$\forall x(\texttt{AsLifelikeAs}(x,e) \to \neg\texttt{MyDream}(x)), \tag{6.4}$$

[18] By contrast, $K(p \wedge q) \to Kq$ is valid according to Yablo's view, since $p \wedge q \geq q$, illustrating that $\varphi \geq \psi$ does not imply $\sigma(\varphi) \geq \sigma(\psi)$, where σ is a uniform substitution function.

[19] Williamson [2000] correctly points out that AC fails when an agent does not grasp the new disjunct ψ, but I have been assuming that our IAL does grasp it, and Williamson does not argue against such a restricted version of AC.

which is of course logically equivalent to (6.3). Yablo [2012a] concludes that "Apparently it is easier to know about dreams that that they are not this lifelike than it is to know about experiences this lifelike that they are not dreams" (12).

I do not have the reported intuition when considering (6.3) and (6.4). Since I assume the students were exposed to natural language utterances rather than (6.3) and (6.4), one would have to think carefully about possible context change, default reasoning, etc., before concluding that the situation is best represented as one in which students think it is harder to know (6.4) than the classically equivalent (6.3).

According to Yablo, the reason that (6.3) is easier to know than (6.4) is that they differ in subject matter (the first is supposed to be about my dreams, whereas the second is supposed to be about my phenomenal states individuated qualitatively). However, while Yablo's example involves equivalence in first-order logic, it follows from his formal theory of subject matter for propositional logic in Definition 6.2 that logically equivalent propositional formulas have the *same* subject matter.

Fact 6.1 (Equivalence and Subject Matter)**.** If φ and ψ are classically equivalent propositional formulas, then $\varphi \geq \psi$ and $\psi \geq \varphi$.

Proof. Assume all classical models of φ are classical models of ψ and vice versa. Suppose v is a partial model of φ, so for all classical valuations v extending v, $v \vDash \varphi$. By the assumption, it follows that for all classical valuations v extending v, $v \vDash \psi$, so v is also a partial model of ψ. Hence all partial models of φ are partial models of ψ, and obviously vice versa, which implies that all minimal models of φ are minimal models of ψ and vice versa. The same reasoning shows that all minimal models of $\neg\varphi$ are minimal models of $\neg\psi$. It follows by Definition 6.2 that $\varphi \geq \psi$ and $\psi \geq \varphi$. □

It follows from Fact 6.1 that Yablo's theory of subject matter and its relation to closure does not undermine closure under propositional logical equivalence. Whether an extension of his theory to first-order logic would undermine closure under first-order logical equivalence is a question that awaits the development of such an extension.

It is worth mentioning a different formalization of the idea of subject matter containment, related to Parry's [1933, 1989] notion of *analytic implication* (also see Anderson and Belnap 1975, §29.6), that does distinguish logical equivalents. Burgess

[2009, §5.2] considers an extension of classical logic that he calls *topic logic*, adding a new binary connective / (among others) to the propositional language. A model for (propositional) topic logic is a pair $\langle v, s\rangle$, where v assigns each sentence letter a truth value and s assigns each sentence letter a subset of some set T, thought of as a "set of topics." We extend s to a function $\hat{s}$ on arbitrary formulas as follows:

$$\begin{aligned} \hat{s}(p) &= s(p) \\ \hat{s}(\neg\varphi) &= \hat{s}(\varphi) \\ \hat{s}(\varphi \# \psi) &= \hat{s}(\varphi) \cup \hat{s}(\psi) \end{aligned}$$

for any two-place connective #. Hence $\hat{s}(\varphi)$ is the union of the $s(p)$ sets for every sentence letter p occurring in φ. (Cf. Lewis 1988, 155: "The part-of-statements conception [of partial aboutness] is *cumulative*. When we build up statements from their parts, we may gain new subject matters for the resulting statement to be partly about, but we never lose old ones.") The valuation v extends to a valuation $\hat{v}$ for formulas with the classical connectives in the usual way, and for the new connective /, Burgess defines $\hat{v}(\varphi/\psi) = 1$ iff $s(\psi) \subseteq s(\varphi)$, in other words, iff the subject matter of φ contains that of ψ. As an abbreviation, let us write $\varphi \rightarrow / \psi$ for $(\varphi \rightarrow \psi) \wedge (\varphi/\psi)$.

As Burgess observes, if φ and ψ are formulas without the new connective /, then $\varphi \rightarrow / \psi$ is valid (true in all models $\langle v, s\rangle$) iff ψ is a classical consequence of φ and the set of sentence letters occurring in φ contains the set of sentence letters occurring in ψ (in Parry's [1933] terminology, φ is "analytically relevant" to ψ). For φ and ψ without the new connective, it is interesting to compare when $\varphi \rightarrow / \psi$ is valid in topic logic to when $\varphi \geq \psi$ holds according to Yablo's Definition 6.2, especially when we consider Yablo's proposal that closure should hold ($K\varphi \rightarrow K\psi$ should be valid) barring a change in subject matter from φ to ψ. It is easy to see that the two conditions are incomparable in strength. For example, the topic logic condition supports $K(\varphi \wedge \psi) \rightarrow K(\varphi \vee \psi)$ and $K(\varphi \wedge \psi) \rightarrow K\varphi$ for any φ and ψ, since $(\varphi \wedge \psi) \rightarrow / (\varphi \vee \psi)$ and $(\varphi \wedge \psi) \rightarrow / \varphi$ are valid for any φ and ψ, whereas Yablo's condition does not (recall above). On the other hand, the topic logic condition does not support closure under classical equivalence, whereas Yablo's does (Fact 6.1).

As I have indicated, I take the failure of $K(\varphi \wedge \psi) \rightarrow K\varphi$ to tell against the requirement on $K\varphi \rightarrow K\psi$ that $\varphi \geq \psi$. Although the requirement that $\varphi \rightarrow / \psi$ be valid does not have this consequence, it does lead to the rejection of AC, $K\varphi \rightarrow K(\varphi \vee \psi)$, which is too much for me. (I should note that Burgess's discussion of topic logic has nothing to do with epistemic closure.) While I reject the closure step from Kp to $K(p \wedge \neg s)$, I accept the step from Kp to $K(p \vee \neg s)$. In my view, there is an important difference between the two: relative to p, $p \wedge \neg s$ says *more about more* in the sense explained in §6.1.2; whereas relative to p, $p \vee \neg s$ says *less about more* (logically weaker with expanded subject matter).[20] I hold that knowing *more about more* can require more epistemic work. Yablo agrees but adds that knowing *less about more* can require more epistemic work too. In the case of AC, I agree with Dretske, Nozick, and Kripke[21] that knowing φ is a path to knowing $\varphi \vee \psi$. In §6.2.2 - 6.2.3, I consider responses to the issue of (6.1) and (6.2) that maintain AC.

6.2.2 Dretske and the Denial of RE

According to Yablo's view, the problematic principle in the derivation of (13) - (16) is $Kp \rightarrow K(p \vee \neg s)$, whereas by Fact 6.1, $K(p \vee \neg s) \rightarrow K\neg(\neg p \wedge s)$ holds on Yablo's view. I will now consider views according to which Yablo's view gets things backwards. For those who, like Dretske [1970], Nozick [1981], and Kripke [2011], wish to maintain $K\varphi \rightarrow K(\varphi \vee \psi)$, but who, like Dretske and Nozick, reject either (6.1)

[20]Note added in ILLC version: in contrast to Burgess's idea of the set of topics that a proposition is about, Dunn [1976, §6] introduces the idea of the set of topics that a proposition gives *definite information about*. To a given propositional formula φ, Dunn assigns a pair $I(\varphi) = \langle I^+(\varphi), I^-(\varphi)\rangle$, where $I^+(\varphi)$ is the set of topics that φ gives definite information about and $I^-(\varphi)$ is the set of topics that the negation of φ gives definite information about. Given an assignment of such pairs to atomic propositions, Dunn uses the following recursive clauses: $I(\neg\varphi) = \langle I^-(\varphi), I^+(\varphi)\rangle$; $I(\varphi \wedge \psi) = \langle I^+(\varphi) \cup I^+(\psi), I^-(\varphi) \cap I^-(\psi)\rangle$; and $I(\varphi \vee \psi) = \langle I^+(\varphi) \cap I^+(\psi), I^-(\varphi) \cup I^-(\psi)\rangle$. Note that while a disjunction may introduce new topics according to Burgess, it does not give *definite information about* new topics according to Dunn. Thus, if we adopt the view that an epistemic closure step moving from φ to a logical consequence ψ is problematic only if ψ gives *definite information about* new topics—as opposed to just being about new topics—then an epistemic closure step moving from α to $\alpha \vee \beta$ is not problematic, because $I^+(\alpha \vee \beta) \subseteq I^+(\alpha)$, whereas an epistemic closure step from p to $p \wedge \neg s$ may well be problematic, because there is no guarantee that $I^+(p \wedge \neg s) \subseteq I^+(p)$.

[21]Kripke [2011, 202] writes that "I myself believe that for the intuitive concept of knowledge, adding a disjunct ought to preserve knowledge."

or (6.2), there is only one choice: give up the "De Morgan" closure principle $K(\pm\varphi \vee \pm\psi) \rightarrow K\neg(\mp\varphi \wedge \mp\psi)$ and hence closure under logical equivalence. Having given up this instance of single-premise closure, one faces the question of what distinguishes $K(\pm\varphi \vee \pm\psi) \rightarrow K\neg(\mp\varphi \wedge \mp\psi)$ from the principles that Dretske and Nozick accept.

According to Yablo's view, the problem with (6.1) and (6.2) is that their consequents, although logically weaker than their antecedents, introduce new subject matter. In my terminology, they involve knowing less about more. Now I will offer an alternative explanation of why (6.1) and (6.2) seem problematic: their consequents claim knowledge that something is *not* the case, and this *negation* brings with it the idea of **contrast** that I argued fallibilists should not accept in general (recall §5.1.2).[22] In particular, I argued that **contrast** can fail for disjunctions like $p \vee \neg s$; for I agree with Dretske and Kripke that one path to knowing $p \vee \neg s$ is via knowing p, and I agree with fallibilists in general that coming to know p may not require ruling out $(\neg p \wedge s)$-worlds, so I conclude that ruling out $(\neg p \wedge \neg s)$-worlds may be sufficient for knowing $p \vee \neg s$. But can one come to know $\neg(\neg p \wedge s)$ without ruling out $(\neg p \wedge s)$-worlds?

With the negated conjunction, I expect some peoples' intuitions to shift in favor of **contrast**, perhaps because the processing of negations in non-epistemic contexts in natural language involves the construction of contrast classes (see Oaksford and Stenning 1992). There are three ways the explanation might go from here:

1. $K(\pm\varphi \vee \pm\psi) \rightarrow K\neg(\mp\varphi \wedge \mp\psi)$ is valid, but the negation in the consequent triggers the mistaken intuition that **contrast** must hold for $\mathbf{r}(\neg(\mp\varphi \wedge \mp\psi), w)$.

2. $K(\pm\varphi \vee \pm\psi) \rightarrow K\neg(\mp\varphi \wedge \mp\psi)$ is valid for a fixed context, but when an attributor claims that an agent knows that *not-P*, this has a tendency to change the context to one in which **contrast** holds for $\mathbf{r}(P, w)$ (cf. DeRose 1995).

3. $K(\pm\varphi \vee \pm\psi) \rightarrow K\neg(\mp\varphi \wedge \mp\psi)$ is not valid even for a fixed context, because **contrast** may apply to $\mathbf{r}(\neg(\mp\varphi \wedge \mp\psi), w)$ without applying to $\mathbf{r}(\pm\varphi \vee \pm\psi, w)$.

In this section, I will consider the third position, which is the only one available to someone like Dretske who rejects epistemic contextualism and, as I have argued,

[22] I believe Hannah Ginsborg at Berkeley was the first person from whom I heard the suggestion that what might make certain closure principles problematic was the introduction of negation.

$K(\pm\varphi \vee \pm\psi) \rightarrow K\neg(\mp\varphi \wedge \mp\psi)$. The question is: for those who reject the validity $K(\pm\varphi \vee \pm\psi) \rightarrow K\neg(\mp\varphi \wedge \mp\psi)$, what other closure principles can they accept?

One approach to answering this question is to modify the rule RM, with which we can obtain $K\varphi \rightarrow K\psi$ whenever $\varphi \rightarrow \psi$ is derivable in classical logic, to a weaker rule RM_{S}, with which we can obtain $K\varphi \rightarrow K\psi$ whenever $\varphi \rightarrow \psi$ is derivable in a system S that is weaker than classical logic.[23] For example, from the hypothesis that if (6.1) and (6.2) are not valid, then their non-validity should be explained by the connection between negation and **contrast**, we are lead to another hypothesis: that a sufficient (but not necessary) condition for $K\varphi \rightarrow K\psi$ to be an innocuous single-premise closure principle (at least for IALs) is that $\varphi \rightarrow \psi$ be a valid principle in the *positive fragment* of our language, without negation. In the list of axioms below, together with the rule of modus ponens, axioms 1 - 8 (Hilbert's *positive propositional calculus* or *positive logic*) axiomatize the positive fragment of intuitionistic propositional logic, while 1 - 9 axiomatize the positive fragment of classical propositional logic [Carnielli et al., 2007, Middelburg, 2011], where $\rightarrow$ is now a primitive symbol:

1. $\varphi \rightarrow (\psi \rightarrow \varphi)$
2. $(\varphi \rightarrow (\psi \rightarrow \chi)) \rightarrow ((\varphi \rightarrow \psi) \rightarrow (\varphi \rightarrow \chi))$
3. $(\varphi \wedge \psi) \rightarrow \varphi$
4. $(\varphi \wedge \psi) \rightarrow \psi$
5. $\varphi \rightarrow (\varphi \vee \psi)$
6. $\psi \rightarrow (\varphi \vee \psi)$
7. $\varphi \rightarrow (\psi \rightarrow (\varphi \wedge \psi))$
8. $(\varphi \rightarrow \psi) \rightarrow ((\chi \rightarrow \psi) \rightarrow ((\varphi \vee \chi) \rightarrow \psi))$
9. $((\varphi \rightarrow \psi) \rightarrow \varphi) \rightarrow \varphi$

[23] The corresponding constraint on the syntactic r function from §5.2.4 would be:

cover$_{\mathsf{S}}$ if $\vdash_{\mathsf{S}} \varphi \rightarrow \psi$, then $\forall B \in r(\varphi, w)\ \exists A \in r(\psi, w)$: $A \subseteq B$.

10. modus ponens.

As Church [1995/1956] explains, positive logic was "designed to embody the part of propositional calculus which may be said to be independent in some sense of the existence of negation" (140). Now if we bring negation back into our language, the question arises as to which principles involving negation we should add to 1 - 9 to obtain a system such that if $\varphi \to \psi$ is derivable, then $K\varphi \to K\psi$ is an innocuous closure principle. According to the Dretskean view we are considering, we do not want $(\neg\varphi \vee \neg\psi) \to \neg(\varphi \wedge \psi)$ to be derivable. In standard extensions of positive logic, such as minimal logic[24] and intuitionistic logic,[25] although not all of the De Morgan laws are derivable, $(\neg\varphi \vee \neg\psi) \to \neg(\varphi \wedge \psi)$ is; on the other hand, these systems do not derive the right-to-left direction of $\varphi \leftrightarrow \neg\neg\varphi$, which the Dretskean may want in full. If so, the Dretskean must extend positive logic in a different direction.

However, there is a subtlety here. Suppose that instead of adopting a rule like RM_S, as described above, one adopts a closure principle $K\varphi \to K\psi$ for each axiom $\varphi \to \psi$ listed above, including those for minimal logic in note 24. One might think that this is equivalent to adopting RM_S for the same system. However, the two approaches are not equivalent for a Dretskean who rejects K. To see this, consider the following derivation using minimal logic with RM_S:

(17)	$(m \wedge d) \to m$	axiom
(18)	$((m \wedge d) \to m) \to (\neg m \to \neg(m \wedge d))$	axiom
(19)	$\neg m \to \neg(m \wedge d)$	(17), (18), modus ponens
(20)	$K\neg m \to K\neg(m \wedge d)$	(19), RM_S

Now suppose that instead of RM_S, we have a closure principle $K\varphi \to K\psi$ for each axiom $\varphi \to \psi$, and consider the following derivation:

(21)	$K(m \wedge d) \to Km$	axiom

[24] Obtained by adding $(\varphi \to \psi) \to (\neg\psi \to \neg\varphi)$ and $\varphi \to \neg\neg\varphi$ to positive logic.
[25] Obtained by adding $(\varphi \to \neg\psi) \to (\psi \to \neg\varphi)$ and $\neg\varphi \to (\varphi \to \psi)$ to positive logic.

(22) $K((m \wedge d) \rightarrow m) \rightarrow K(\neg m \rightarrow \neg(m \wedge d))$ axiom

(23) $K(\neg m \rightarrow \neg(m \wedge d))$ (21), (22), modus ponens

If we had the K axiom,[26] then we could extend the derivation as follows:

(24) $K(\neg m \rightarrow \neg(m \wedge d))) \rightarrow (K\neg m \rightarrow K\neg(m \wedge d))$ K axiom

(25) $K\neg m \rightarrow K\neg(m \wedge d)$ (23), (24), modus ponens,

obtaining the allegedly problematic (6.2). However, if like Dretske we do not have the K axiom, then we cannot always epistemically "internalize" proofs of theorems in system S, which is why the approach with RM_{S} and the approach with a closure principle for each axiom are not equivalent (I leave a proper proof of the inequivalence to the reader). Note that if we adopt the approach with a closure principle for each axiom, then propositional axiom systems that are equivalent given modus ponens will not necessarily give rise to equivalent epistemic logics without the K axiom.

Finally, I will mention one other approach that is consistent with the denial of RE. Presented with the arguments for the Multipath Picture in Chapter 5, a number of people have asked whether $r(\varphi, w)$ (where r is the syntactic version of $\mathbf{r}$ discussed in §5.2.4) could be defined by recursion on the structure of φ. My answer is that we can at least put natural constraints on r that are of a recursive character.[27] For example, we could adopt the following constraints:

($\vee$-paths) $\forall B \in r(\varphi, w) \cup r(\psi, w)\ \exists A \in r(\varphi \vee \psi, w)\colon\ A \subseteq B$;

($\wedge$-paths) $\forall B \in r(\varphi \wedge \psi, w)\ \exists A \in r(\varphi, w)\ \exists A' \in r(\psi, w)\colon\ A \cup A' \subseteq B$,

corresponding to $(K\varphi \vee K\psi) \rightarrow K(\varphi \vee \psi)$ and $K(\varphi \wedge \psi) \rightarrow (K\varphi \wedge K\psi)$, respectively. However, *defining* r by recursion is a different matter. Those who accept the principle $(K\varphi \wedge K\psi) \leftrightarrow K(\varphi \wedge \psi)$ could define

$$r(\varphi \wedge \psi, w) = \{A \cup A' \mid A \in r(\varphi, w) \text{ and } A' \in r(\psi, w)\},$$

[26] Here written as $K(\varphi \rightarrow \psi) \rightarrow (K\varphi \rightarrow K\psi)$ instead of $(K\varphi \wedge K(\varphi \rightarrow \psi)) \rightarrow K\psi$.

[27] Note added in ILLC version: for a more definite answer, see the recursive construction in the Multipath Theorem of Holliday 2013b.

but to give an equation for $r(\varphi \vee \psi, w)$, we seem to need more than just $r(\varphi, w)$ and $r(\varphi, w)$, since we need the path to knowing $\varphi \vee \psi$ without knowing either disjunct.[28] Moreover, negation is again a problem: it is not clear how to define $r(\neg\varphi, w)$ in terms of $r(\varphi, w)$, although if we take $r(\neg p, w)$ as given, then we could define

$$r(\neg\neg\varphi, w) = r(\varphi, w);$$

$$r(\neg(\varphi \wedge \psi), w) = r(\neg\varphi \vee \neg\psi, w);$$

$$r(\neg(\varphi \vee \psi), w) = r(\neg\varphi \wedge \neg\psi, w).$$[29]

Of course, since I have argued that Dretske must reject $K(\neg\varphi \vee \neg\psi) \rightarrow K\neg(\varphi \wedge \psi)$, he would at least have to weaken the second equation to

$$\forall B \in r(\neg(\varphi \wedge \psi), w)\ \exists A \in r(\neg\varphi \vee \neg\psi, w):\ A \subseteq B.$$

Finally, observe that if we adopt the recursive constraints rather than a recursive definition, then we have not yet ensured such trivial "closure principles" as $K(\varphi \vee \psi) \rightarrow K(\psi \vee \varphi)$. Here we might adopt the view that $K(\varphi \vee \psi)$ and $K(\psi \vee \varphi)$ are mere syntactic variants that do not represent different knowledge ascriptions at all, so that $K(\varphi \vee \psi) \rightarrow K(\psi \vee \varphi)$ should be thought of as a re-writing principle rather than a closure principle. More generally, we might define a syntactic variant relation $\approx$ on formulas, e.g., with at least the following conditions (cf. Levesque 1984, §5):

$$(\varphi \wedge \psi) \approx (\psi \wedge \varphi) \qquad (\varphi \wedge \varphi) \approx \varphi$$

$$(\varphi \vee \psi) \approx (\psi \vee \varphi) \qquad (\varphi \vee \varphi) \approx \varphi$$

$$(\varphi \wedge (\psi \wedge \chi)) \approx ((\varphi \wedge \psi) \wedge \chi) \qquad \varphi \approx \neg\neg\varphi$$

$$(\varphi \vee (\psi \vee \chi)) \approx ((\varphi \vee \psi) \vee \chi)$$

$$(\varphi \leftrightarrow \psi) \approx (\psi \leftrightarrow \varphi) \qquad \text{if } \alpha \approx \beta, \text{ then } \varphi(\alpha/p) \approx \varphi(\beta/p).$$

One could then adopt the constraint that $r(\varphi, w) = r(\psi, w)$ whenever $\varphi \approx \psi$.

[28] Here one might look to an ordering of worlds, taking the set of closest $\neg(\varphi \vee \psi)$-worlds.

[29] Recall the discussion of the r_{-} function in §5.2.4.

All of the ideas suggested in this section could be explore in much greater depth. However, I will conclude this section by stressing the following crucial facts: first, all of the constraints on r that we have considered in this section (including in note 23), as well as any that someone who accepts less than full single-premise closure would be willing to consider, are weaker than the *cover* constraint discussed in §5.2.4; second, in §5.2.4 the *cover* constraint was shown to be consistent with all of the other postulates in the Multipath Picture, including the *noVK* postulate that eliminated the Problem of Vacuous Knowledge; it follows that any constraints weaker than *cover* are also consistent with the other postulates and *noVK*. In other words, by adopting the Multipath Picture, fallibilists who accept less than full single-premise closure can rest assured that whatever system they settle upon, they will avoid the Problem of Vacuous Knowledge. This is a significant result, since we saw in §4.B.1 that by adding only $K\varphi \rightarrow K(\varphi \vee \psi)$ or $K(\varphi \wedge \psi) \rightarrow K\varphi$ and $K\varphi \leftrightarrow K((\varphi \vee \psi) \wedge \varphi)$, let alone full single-premise closure, to Fallibilism 1.0, we brought on the Problem of Vacuous Knowledge. Hence we had to choose between the Problem of Vacuous Knowledge or giving up one of those three principles, exemplifying the Problem of Containment. By contrast, all of the approaches considered in this section secure those principles, and given the Multipath Picture in §5.2.4, they can do so without vacuous knowledge.

6.2.3 Roush and the Defense of RM

Recall our three options when faced with fellow fallibilists like Dretske and Nozick who reject (6.1) and (6.2):

1. Deny AC, $K\varphi \rightarrow K(\varphi \vee \psi)$, even for IALs.

2. Deny the "De Morgan" closure principle $K(\pm\varphi \vee \pm\psi) \rightarrow K\neg(\mp\varphi \wedge \mp\psi)$ and hence closure under logical equivalence, RE, even for IALs.

3. Defend single-premise logical closure, RM, and hence (6.1) and (6.2), at least for IALs.

In this final section, I will consider the third option, according to which Dretske and Nozick were wrong to reject (6.1) and (6.2),

$$Kp \rightarrow K\neg(\neg p \wedge s) \text{ and}$$
$$K\neg p \rightarrow K\neg(p \wedge s).$$

In the previous chapters, I argued that fallibilists who wish to maintain full multi-premise closure are lead to either the Problem of Vacuous Knowledge or the Problem of Knowledge Inflation, whereas fallibilists who maintain single-premise logical closure are not. What reasons are there to reject the single-premise (6.1) and (6.2)? Nozick notes that (6.1) fails according to his theory, but he offers no independent argument against it. Since we have read between the lines to see that Dretske must reject (6.2), we do not have an argument from him either. However, the source of worries about (6.1) and (6.2) is clear: it is the idea that they are dangerous in the hands of skeptics. To see whether this is so, we must first review the skeptic's tactics.

Standard Skeptical Arguments

Let p be the proposition that *b is a Gadwall*, and let s be the skeptical hypothesis that *b is a Siberian Grebe*. Read $K\varphi$ as the *third-person* knowledge attribution, "Alice knows that φ." Below I state the standard, multi-premise skeptical argument. As in §6.1.2, each line contains the following information from left to right: line number, formula, justification, set of open assumptions, and my evaluation ($\times$ indicates rejection and in later arguments $\checkmark$ indicates endorsement).

Skeptical Argument I

(26)	$\neg K\neg s$	premise	$\{(26)\}$	granted
(27)	$K(p \rightarrow \neg s)$	premise	$\{(27)\}$	granted
(28)	$(Kp \wedge K(p \rightarrow \neg s)) \rightarrow K\neg s$	K	$\{\}$	$\times$
(29)	$\neg Kp$	(26) - (28), PL	$\{(26), (27)\}$	$\times$

Let me emphasize two points. First, if we substitute $(Kp \wedge K\Box(p \rightarrow \neg s)) \rightarrow K\neg s$ for (27), I reject that step as well. Second, if I am in the role of the fallibilist park ranger in DIALOGUE I, then I will continue to maintain that Alice knows that the bird is a Gadwall, even after the skeptical ornothologist raises in (26) the Siberian Grebe hypothesis that we both agree is false. I do not think, as some contextualists do, that by merely raising this hypothesis, the ornothologist changes the context to one in which the "skeptic wins" and Alice no longer counts as knowing that the bird is a Gadwall. In a case like DIALOGUE I, I stand my fallibilist ground. Yet I am willing to grant the skeptic his first premise in cases where Alice has little background information about Siberian Grebes and has not seen the bird's belly in flight.[30]

Recall that the closure step in ARGUMENT I is an example of *multi-premise* logical closure because $\neg s$ is not a logical consequence of the single premise p, though it is a logical consequence of the set of premises $\{p, p \rightarrow \neg s\}$. Here we meet the objection: while $\neg s$ is not a logical consequence of p, the skeptic can just run his argument with $\neg(\neg p \wedge s)$ instead, since that *is* a logical consequence of p.[31] But if so, the objector says, then those fallibilists who deny closure against the skeptic must even give up single-premise logical closure, right? Not so fast. Let us analyze the argument:

SKEPTICAL ARGUMENT II

(30)	$\neg K\neg(\neg p \wedge s)$	premise	$\{(30)\}$
(31)	$Kp \rightarrow K\neg(\neg p \wedge s)$	(6.1)	$\{\}$
(32)	$\neg Kp$	(30), (31), PL	$\{(30)\}$

According to Roush [2010], the move from $\neg s$ in (26) to $\neg(\neg p \wedge s)$ in (30) trivializes the skeptic's argument. Before explaining this view, let us consider a skeptical argument whose first premise is blatantly *question-begging* with respect to its conclusion:

[30] Recall that as Dretske sets up the case, the only way to tell apart a Gadwall and a Siberian Grebe is to look at the markings on the belly of the bird when it is in flight.

[31] One might call this the "BIV to *handless* BIV" maneuver, to which I return in §6.2.3.

(33)	$\neg Kp$	premise	{(33)}
(34)	$Kp \rightarrow Kp$	tautology	{}
(35)	$\neg Kp$	(33), (34), PL	{(33)}

Next, observe that the skeptical argument in (36) - (38) is at least as bad as the one in (33) - (35) on the score of question-begging. Intuitively, the skeptic assumes in (36) that the agent has an even greater lack of knowledge than assumed in (33):

(36)	$\neg K(p \vee \neg s)$	premise	{(36)}
(37)	$Kp \rightarrow K(p \vee \neg s)$	AC	{}
(38)	$\neg Kp$	(36), (37), PL	{(36)}

From here, consider the *closure of question-begging under equivalence*: if P is question-begging as a premise for conclusion C, and P is equivalent to P', then P' is question-begging as a premise for C.[32] If this is correct, and if we can assume that $\neg K(p \vee \neg s)$ is equivalent to $\neg K\neg(\neg p \wedge s)$ for IALs, then given that $\neg K(p \vee \neg s)$, like $\neg Kp$, is question-begging as a premise for the conclusion of $\neg Kp$, it follows that $\neg K\neg(\neg p \wedge s)$ is question-begging as a premise for the conclusion of $\neg Kp$ in Argument II.

I take the closure principle for question-begging to be uncontroversial. For I know of no example in which a philosopher has responded to the charge that a premise is question-begging by substituting an admittedly equivalent premise. To avoid the results that (30) is question-begging in Argument II, it seems that one must deny that $\neg K\neg(\neg p \wedge s)$ is equivalent to $\neg K(p \vee \neg s)$ for IALs, which means denying that $K\neg(\neg p \wedge s)$ is equivalent to $K(p \vee \neg s)$ for IALs, which leads one back to §6.2.2.

In the next section, I consider Roush's take on Argument II when the argument schema is instantiated with a special kind of skeptical hypothesis.

[32] We do not need a principle of such generality, but the basic intuition should be clear. I expect that any plausible refinement of the principle would also apply in the case at hand.

Self-Side Skeptical Hypotheses

Roush's [2010] discussion concentrates on what I call *self-side* skeptical hypotheses, as opposed to the *world-side* skeptical hypotheses that I have considered in Examples 1.1 and 1.2. Austin [1946, 158] clearly states the distinction I have in mind:

> ...either my current experiencing or the item currently under consideration (or uncertain which) may be abnormal, *phoney*. Either I myself may be dreaming, or in delirium, or under the influence of mescal, etc.: or else the item may be stuffed, painted, dummy, artificial, trick, freak, toy, assumed, feigned, etc.: or else again there's an uncertainty (it's left open) whether *I* am to blame or *it* is—mirages, mirror images, odd lighting effects, etc.

A skeptical hypothesis according to which the item currently under consideration is abnormal or phoney is a *world-side* skeptical hypothesis, whereas a skeptical hypothesis according to which the agent's experiencing is abnormal or phoney is a *self-side* skeptical hypothesis. For example, let `Matrix` be the self-side skeptical hypothesis that Alice's brain is being stimulated by a computer simulation as in the movie The Matrix. Below is the crudest self-side skeptical argument. I will now omit the justification and open assumptions for each step, trusting the reader to fill them in:

(39) $\neg K \neg \texttt{Matrix}$ granted

(40) $K\texttt{Gadwall} \to K \neg \texttt{Matrix}$ ×

(41) $\neg K\texttt{Gadwall}$ ×

What is the skeptic's argument for (40)? Faced with a fallibilist, the skeptic might try to appeal to closure under known entailment:

(42) $K\Box(\texttt{Gadwall} \to \neg\texttt{Matrix})$ ×

(43) $(K\texttt{Gadwall} \wedge K\Box(\texttt{Gadwall} \to \neg\texttt{Matrix})) \to \neg K \neg \texttt{Matrix}$ false antecedent

(44) $K\texttt{Gadwall} \to K \neg \texttt{Matrix}$ ×

One problem with this argument is that the proposition $\texttt{Gadwall}$ about the external world does *not* entail the negation of the self-side skeptical hypothesis $\texttt{Matrix}$; $\Box(\texttt{Gadwall} \rightarrow \neg\texttt{Matrix})$ is false, so it is not known. For it is compatible with b being a Gadwall that Alice's brain is being stimulated by a computer simulation. G.E. Moore's famous Duke of Devonshire story makes this point about dreaming; Stroud [1984, 25-29] emphasizes it; and Roush [2010] argues that it deflates closure-based arguments for skepticism that use what I call self-side skeptical hypotheses.

The skeptic might claim that although $K\Box(\texttt{Gadwall} \rightarrow \neg\texttt{Matrix})$ is false, we can assume that $K(\texttt{Gadwall} \rightarrow \neg\texttt{Matrix})$, i.e, $K(\neg\texttt{Gadwall} \vee \neg\texttt{Matrix})$ is true, so the argument can be reformulated using closure under known material implication instead of closure under known entailment. In that case, I would reject the argument at the point where it appeals to closure under known material implication.

Although obvious, it is worth observing that if we replace the proposition that *b is a Gadwall* with the proposition that *Alice has hands*, it does not improve the argument (in the movie The Matrix, the subjects of brain stimulation all have hands):

(45) $\neg K\neg\texttt{Matrix}$ granted

(46) $K\Box(\texttt{hands} \rightarrow \neg\texttt{Matrix})$ ×

(47) $(K\texttt{hands} \wedge K\Box(\texttt{hands} \rightarrow \neg\texttt{Matrix})) \rightarrow \neg K\neg\texttt{Matrix}$ false antecedent

(48) $\neg K\texttt{hands}$ ×

At this point, the skeptic may say that what he meant all along by his skeptical hypothesis was not $\texttt{Matrix}$ but rather $(\neg\texttt{hands} \wedge \texttt{Matrix})$, arguing as follows:

(49) $\neg K\neg(\neg\texttt{hands} \wedge \texttt{Matrix})$ ×

(50) $\texttt{hands} \rightarrow \neg(\neg\texttt{hands} \wedge \texttt{Matrix})$ ✓

(51) $K\texttt{hands} \rightarrow K\neg(\neg\texttt{hands} \wedge \texttt{Matrix})$?

(52) $\neg K\texttt{hands}$ ×

Now we have a logically valid implication in (50), but as Roush [2010, 245] argues:

> However, it is not enough that there be an implication. It must be an implication from something we think we do know to something we pretty clearly do not, in order to set us up for a modus tollens. What is wrong with this particular patch is that weakening the conclusion to "I am not a handless brain in a vat" trivializes it for this purpose. If we assume I know that I have a hand, then we should not have the slightest hesitation to credit me with knowledge that I am not a handless brain in a vat.
>
> No appeal to the closure principle is needed to support this conclusion. The claim is independently obvious because that you are not a handless brain in a vat is just not much to know. If we know that someone has hands then it follows that she is not a handless person with high blood pressure, or a handless victim of child abuse, but this would not give us any assurance that she need not go to a doctor for these conditions .... If I know that I have hands, then in virtue of that I know I am not a handless anything. The implication is achieved in the skeptical argument, but only by letting the issue of brains in vats swing free of it.

To underscore Roush's last point, when we claim that an agent knows $\neg(\neg\texttt{hands} \wedge \texttt{Matrix})$, we are *not* claiming that the agent knows the logically stronger $\neg\texttt{Matrix}$. We must not let the skeptic blind us to the logical fact that $\neg\texttt{Matrix}$ does not follow from $\neg(\neg\texttt{hands}\wedge\texttt{Matrix})$, so his arguments that $\neg\texttt{Matrix}$ is difficult or impossible to know do not directly show that the weaker $\neg(\neg\texttt{hands} \wedge \texttt{Matrix})$ is difficult or impossible to know. This point raises the question of to what extent self-side skepticism of the "handless BIV" variety trades on luring people into incorrect intuitions about the distribution of negation over conjunction. I don't know, but we shouldn't fall for it.[33]

Roush [2010] defends the position stated in the quote above at length, but I will not repeat her arguments here. Instead, in the next subsection I will show how her basic point applies to world-side skeptical hypotheses of a special kind.

[33]Note added in ILLC version: As Wright [forthcoming] remarks, "Maybe we are confused by the operation of some kind of implicature here: maybe saying, or thinking, "It is not the case that those animals are cleverly disguised mules" somehow implicates, in any context of a certain (normal) kind, that "Those animals have not been cleverly disguised". But anyway, it doesn't entail it: not-(P&Q), dear reader, does not entail not-Q!" (§IV).

Twisted Skeptical Hypotheses

As we have seen, to raise doubts about whether Alice knows that the bird is a Gadwall, one hypothesis that the skeptic can raise is that the bird is instead a *Siberian Grebe*. Another hypothesis that the skeptic can raise is that the bird is instead a *Mallard disguised to look just like a Gadwall*. What distinguishes the second from the first is that the second takes a reasonable alternative that one may well (need to) rule out on the way to knowing that the bird is a Gadwall, the Mallard alternative, and then puts a skeptical twist on it with the idea of disguise. By contrast, the Siberian Grebe alternative is skeptical from the start. I will call the skeptical hypothesis that the bird is a Mallard disguised to look just like a Gadwall a *twisted* skeptical hypothesis, and I will call the hypothesis that the bird is a Siberian Grebe a *direct* skeptical hypothesis.

According to any reasonable fallibilist view, it may well be that Alice not only knows that b is a Gadwall, but also knows that b is not a Mallard. As in DIALOGUE I, assume that the fallibilist and the skeptic are completely agreed that the animal is a Gadwall and not a Mallard, there is no funny business going on, etc. While female Gadwalls and Mallards have similar plumage, suppose Alice has correctly observed that b does not have the characteristic dark orange-edged bill of the female Mallard. Still, the skeptic might try to argue that Alice does not know that b is not a Mallard by using a twisted skeptical hypothesis. Let `Mallard` stand for *b is a Mallard*, and let `disguised` stand for *b is disguised to look just like a Gadwall*:

SKEPTICAL ARGUMENT III

(17) $\neg K \neg (\texttt{Mallard} \wedge \texttt{disguised})$ ×

(18) $\neg\texttt{Mallard} \rightarrow \neg(\texttt{Mallard} \wedge \texttt{disguised})$ ✓

(19) $K\neg\texttt{Mallard} \rightarrow K\neg(\texttt{Mallard} \wedge \texttt{disguised})$?

(20) $\neg K \neg\texttt{Mallard}$ ×

Of course, this is just a special case of ARGUMENT II. Hence Roush's line applies here as well: if Alice knows it's not the case that b is a Mallard, then Alice knows the

logically weaker proposition that: it's not the case that [b is a Mallard **and** *anything*], assuming she believes it. Perhaps filling in the 'anything' with a bizarre proposition will change some attributors' conversational context, badgering Alice about it may shake her confidence, etc., but the claim is that when she knows it's not the case that b is a Mallard and believes it's not the case that [b is a Mallard **and** *anything*], then she also knows the weaker proposition. Note that "this is not a claim about what it would be appropriate to say, what the person himself thinks he knows or would say he knows. It is a question, simply, of what he knows" [Dretske, 1971, 1009 - 1010].

One may object that if a fallibilist is willing to hold that $K\neg(\texttt{Mallard}\wedge\texttt{disguised})$, why is the fallibilist not also willing to hold that $K\neg\texttt{SiberianGrebe}$? Why is there a difference between knowing the negations of twisted and direct skeptical hypotheses? The answer is that there are recognized ways, well-established by the practices of birdwatchers, of checking that something thought to be a Gadwall is not a Mallard, including checking for the dark orange-edged bill. Since Alice has performed the necessary checks, she knows $\neg\texttt{Mallard}$, which gives us an explanation of how she knows the logically weaker $\neg(\texttt{Mallard}\wedge\texttt{anything})$, assuming she believes it. Now there are also recognized ways of checking that something thought to be a Gadwall is not a Siberian Grebe, principally checking the color of the belly—white or red—of the bird in flight. However, Alice *has not performed these checks*, so we have no good explanation of how she knows $\neg\texttt{SiberianGrebe}$ (assuming as before that she does not have much background information about Siberian Grebes). At this point others will resort to vacuous knowledge or knowledge inflation, but I will not.

To claim Alice knows $\neg(\texttt{Mallard}\wedge\texttt{disguised})$ is not to ascribe vacuous knowledge to her. In the case we are imagining, Alice has come to know $\neg(\texttt{Mallard}\wedge\texttt{disguised})$ through the empirical work of ruling out $\texttt{Mallard}$-possibilities in order to know $\neg\texttt{Mallard}$. By contrast, the Problem of Vacuous Knowledge arises in Fallibilism 1.0 because an agent can supposedly know $\neg(\texttt{Mallard}\wedge\texttt{disguised})$ without ruling out *any* possibilities, let alone the $\texttt{Mallard}$-possibilities necessary to know $\neg\texttt{Mallard}$.

To claim that Alice knows $\neg(\texttt{Mallard}\wedge\texttt{disguised})$ is also not to endorse knowledge inflation. To endorse knowledge inflation is roughly to claim that there are propositions P and Q such that (i) P implies Q, (ii) coming to know Q by empirical

investigation would require ruling out possibilities that are not required to know P by empirical investigation, but (iii) if someone who knows P—on the basis of empirical investigation insufficient to know Q—goes on to deduce Q from P, then she knows Q. By contrast, on the view of single-premise closure we have been considering, the reason it holds in this case is that coming to know $\neg(\texttt{Mallard} \wedge \texttt{disguised})$ does not require more empirical investigation that coming to know the stronger $\neg\texttt{Mallard}$.

This is the heart of the matter: does knowing the weaker $\neg(\texttt{Mallard} \wedge \texttt{disguised})$ require more empirical work than knowing the stronger $\neg\texttt{Mallard}$? We considered such a view, which rejects $K\neg p \rightarrow K\neg(p \wedge q)$, in §6.2.2, showing that it is consistent with maintaining many other closure principles in the Multipath Picture.

Another way to go is contextualist, holding that $K\neg p \rightarrow K\neg(p \wedge q)$ is valid with respect to a fixed context, but bringing up the issue of `disguised` in conversation may shift our context to one in which Alice must do more empirical work in order to count as knowing both $\neg\texttt{Mallard}$ and $\neg(\texttt{Mallard} \wedge \texttt{disguised})$, relative to the empirical work she must do in order to count as knowing them in our original context. In §4.1, I argued against appealing to contextualism in order to defend fixed-context multi-premise closure principles that commit us to vacuous knowledge. However, as explained above, in the Multipath Picture $K\neg p \rightarrow K\neg(p \wedge q)$ does not commit us to vacuous knowledge, so the contextualist may have a tenable position in this case.

As we have seen, Roush defends single-premise logical closure without appeal to the semantic thesis of contextualism. Roush's explanation is instead partly pragmatic. While it is a logical error to distribute the negation in $\neg(\texttt{Mallard} \wedge \texttt{disguised})$ to obtain $\neg\texttt{disguised}$, Roush [2010, 246] points outs that uttering the English translation of $\neg(\texttt{Mallard} \wedge \texttt{disguised})$ may produce the conversational implicature that $\neg\texttt{disguised}$.[34] If someone says, "Alice knows the bird is not a Mallard" and then adds that "she knows the bird is not a Mallard disguised to look like a Gadwall," this seems to carry a strong conversational implicature that Alice knows the bird is not disguised to look like a Gadwall.[35] However, this implicature can be cancelled by

[34]Note added in ILLC version: recall the quote from Wright in note 33.

[35]Another complication is that when we express the knowledge attributions in English, we tend to use predicate negation (i.e., "the bird is not a Mallard ..."), whereas the epistemic closure principles are stated with sentential negation (i.e., "it is not the case that the bird is a Mallard and ...").

adding "but she doesn't know that the bird is not disguised to look like a Gadwall."[36] All of this reinforces the sense that the pragmatics of negation may be important for knowledge attributions, as suggested in §6.2.2, a topic that deserves further study.

I will not decide between these views here. For one thing, they are not mutually exclusive. For another, each is compatible with the Multipath Picture of Knowledge.

6.3 Conclusion

In this chapter, I have discussed objections according to which accepting single-premise logical closure while denying multi-premise logical closure either admits too much closure or not enough. But the goal, I think, should be to get good lower- and upper-bounds on the line between unproblematic and problematic closure, even if the exact line is a subtle matter of dispute, as witnessed by the analysis of (6.1) and (6.2) in §6.2.2. The results for the Multipath Picture of Knowledge in Chapter 5 show that any lower-bound involving instances of single-premise logical closure can be consistently combined with the rejection of Vacuous Knowledge and Knowledge Inflation, while closure under known (strict) (bi)-implication cannot, leading me to conclude that the appropriate upper-bound is lower than full multi-premise closure.

However, the exact line may not be characterizable in formal terms at all. For example, I have argued that fallibilists should not accept closure under known implication, $(K\varphi \wedge K(\varphi \to \psi)) \to K\psi$, as a general schema for *every* φ and ψ. Are fallibilists who take this line committed to claiming that an IAL can, e.g., know that something is *red* without knowing that it is *colored*? The answer is: no, we are not, and we should say nothing of the sort, unless we can think of a situation in which knowing that something is colored plausibly requires ruling out possibilities that one need not rule out in order to know that it is red. By denying that $(K\varphi \wedge K(\varphi \to \psi)) \to K\psi$ is a valid schema, true for all substitutions for φ and ψ in all situations, we are not committed to claiming that an instance such as $(Kred \wedge K(red \to colored)) \to Kcolored$

[36] To see this clearly, change the story so that Alice has disguised what she knows to be a Pintail as a Gadwall. Then we can say, "Alice knows the bird is not a Mallard" and add that "she knows the bird is not a Mallard disguised to look like a Gadwall—but she doesn't know the bird is not disguised to look like a Gadwall," because she knows that the bird *is* disguised to look like a Gadwall.

can actually be false. Of course, in the formal frameworks of Chapters 2 - 5, the instance $(Kred \wedge K(red \rightarrow colored)) \rightarrow Kcolored$ can be *false in some model*, but this is because the model class is not restricted by taking into account anything about the non-logical content of or relation between *red* and *colored*. If we restricted to models in which *red* and *colored* are interpreted in a certain way and in which the sets of counter-possibilities for these propositions are constrained relative to each other in a certain way, then there would be no falsifying model for the principle.

Following Dretske [2005], one could propose that knowledge is closed under known "lightweight" implications, which do not introduce additional counter-possibilities—counter-possibilities that one must rule out in order to know the consequent ψ but not to know the antecedent φ. For it is only the "heavyweight" implications, leading us from the mundane to the radically skeptical, that force some fallibilists to accept Vacuous Knowledge or Knowledge Inflation in order to maintain closure under known implication. The concept of lightweight implications seems to cover the implication from *it is red* to *it is colored* but not from ordinary propositions to the denial of skeptical hypotheses, as desired. But can we rigorously characterize the class of all φ and ψ that are permissible in the Dretskean principle? Perhaps not. Some may take this as supporting the view that closure under known implication is valid for all φ and ψ after all. In my view, to accept a schema some of whose instances force fallibilists into serious problems, for lack of a certain kind of characterization of what all the problematic instances have in common, would be a serious philosophical mistake. The alternative path, accepting that we must give up full closure, does not seem to me a radical one. Instead, it is the result of following fallibilism where it leads.

6.A The Problem of Factivity

Here is the relevant version of the factivity problem referenced in §6.1.1. Suppose the contextualist finds herself in a context $\mathcal{S}$ in which the skeptic's conversational maneuvers have installed epistemic standards relative to which she does not know some ordinary propositions. The contextualist might like to respond to the skeptic by claiming that she still counts as knowing those ordinary propositions relative to

a context $\mathcal{O}$ with less demanding, ordinary standards. However, by making such a claim in context $\mathcal{S}$, she would be claiming something that is *impossible* for her to know relative to $\mathcal{S}$, as a simple derivation shows. Let $K_{\mathcal{C}}p$ indicate that the agent knows p relative to context $\mathcal{C}$. Hence our initial assumption was $\neg K_{\mathcal{S}}p$. Suppose for reductio ad absurdum that $K_{\mathcal{S}}K_{\mathcal{O}}p$. Relative to any context, knowledge is factive, so $K_{\mathcal{O}}p \rightarrow p$, and we can assume that the contextualist knows this relative to any context, so $K_{\mathcal{S}}(K_{\mathcal{O}}p \rightarrow p)$. Finally, following standard contextualism, we assume closure under known implication relative to any fixed context, $(K_{\mathcal{C}}\varphi \wedge K_{\mathcal{C}}(\varphi \rightarrow \psi)) \rightarrow K_{\mathcal{C}}\psi$, an instance of which is $(K_{\mathcal{S}}K_{\mathcal{O}}p \wedge K_{\mathcal{S}}(K_{\mathcal{O}}p \rightarrow p)) \rightarrow K_{\mathcal{S}}p$. Putting it all together, we derive $K_{\mathcal{S}}p$ and $\neg K_{\mathcal{S}}p$, a contradiction. Hence $K_{\mathcal{S}}K_{\mathcal{O}}p$ is impossible.

Moreover, it is plausible that the contextualist can follow this derivation and come to know that $K_{\mathcal{S}}K_{\mathcal{O}}p$ is impossible. In that case, a weak norm of assertion, namely that you should not assert something that you know to be impossible for you to know relative to what would be the context of your assertion, would prohibit the contextualist from responding to the skeptic as described above.

6.B Subjunctivism and Equivalence

Recall from §6.1.2 Hawthorne's point that "the counterfactual considerations that Dretske and Nozick adduce to divorce the epistemic status of some p from its a priori consequences do not similarly divorce p from its a priori equivalents" (39-40).

The reason, I take it, is that a priori equivalents p and q will be true in exactly the same metaphysically accessible worlds ($\Box(p \leftrightarrow q)$), and if we also assume that they are *believed* by the agent in exactly the same metaphysically accessible worlds ($\Box(Bp \leftrightarrow Bq)$), then any counterfactuals $\varphi(p) \mathrel{\Box\!\!\rightarrow} \psi(p)$ and $\varphi(q) \mathrel{\Box\!\!\rightarrow} \psi(q)$ (where φ and ψ may only contain extensional logical operators and the belief operator B) that differ only with respect to uniform substitution of p and q will have the same truth value (evaluated at the same world in the same context), according to standard semantics for counterfactuals [Stalnaker, 1968, Lewis, 1973]. According to *subjunctivism*, whether one knows p or knows q depends on whether some such counterfactual conditions hold, and by the previous observation, the conditions hold for p iff they hold for q. Given

this reasoning, I agree with Hawthorne, against the claims of Adams et al. [2012], that subjunctivists need a new story if they wish to reject Hawthorne's principle EP.

In their attempt to show that subjunctivists are not committed to EP, Adams et al. assume (without comment) the controversial view that counterfactuals whose antecedents are true in exactly the same metaphysically accessible worlds can differ in truth value. They write: "[I]f it were not the case that x is a zebra, then x would not be a painted mule .... But if it were not the case that x is both a zebra and not a painted mule, ... then x might be a painted mule." Assuming, as usual (see Lewis 1973, §1.1, §1.5), that the might-counterfactual implies the negation of the correspond would-counterfactual with the consequent negated, i.e., $\varphi \diamond\!\!\rightarrow \psi$ implies $\neg(\varphi \mathrel{\square\!\!\rightarrow} \neg\psi)$, the quoted passage commits Adams et al. to the controversial view in question.

The full quote from Adams et al. suggests why they may be lead to their controversial assumption: "But if it were not the case that x is both a zebra and not a painted mule, *i.e., if it were the case that x is a either a non-zebra or a painted mule*, then x might be a painted mule" [emphasis added]. It has been much-discussed in the literature on counterfactuals that a counterfactual with a *disjunctive* antecedent can appear to differ in truth value from a counterfactual (with the same consequent) whose *non-disjunctive* antecedent is true in exactly the same metaphysically accessible worlds as (or is even logically equivalent to) the disjunctive antecedent. Some take this appearance as reality (e.g., Nute 1975, 1978, although Nute 1980 offers an alternative, pragmatic explanation). Others explain it away by distinguishing between counterfactual sentences in natural language with apparently disjunctive antecedents and the true logical form of such sentences (see Lewis 1977 and references therein).

In any case, the intuition of Adams et al. that "if it were not the case that x is a zebra, then x would not be a painted mule" can differ in truth value from "if it were the case that x is either a non-zebra or a painted mule, then x would not be a painted mule" reflects a well-known phenomenon involving disjunctive antecedents. If taken at face value, however, such intuitions would call into question the equivalence that Adams et al. assume (with their "i.e.") between the counterfactual with the disjunctive antecedent and the counterfactual whose antecedent is an equivalent negated conjunction. Then even if they were to argue that Hawthorne's EP fails with

respect to disjunctive propositions in virtue of the behavior of counterfactuals with disjunctive antecedents, what would they say about Hawthorne's argument in (1) - (6)? The authors need to address these delicate issues involving counterfactuals.

Bibliography

Fred Adams, John A. Barker, and Julia Figurelli. Towards closure on closure. *Synthese*, 188(2):179–196, 2012.

Martin Allen. Complexity results for logics of local reasoning and inconsistent belief. In Ron van der Meyden, editor, *Proceedings of the Tenth Conference of Theoretical Aspects of Rationality and Knowledge (TARK X)*, pages 92–108. National University of Singapore, 2005.

Marc Alspector-Kelly. Why safety doesn't save closure. *Synthese*, 183(2):127–142, 2011.

Alan Ross Anderson and Nuel D. Belnap. *Entailment: The Logic of Relevance and Necessity*. Princeton University Press, Princeton, 1975.

Carlos Areces and Balder ten Cate. Hybrid Logics. In Patrick Blackburn, Johan van Benthem, and Frank Wolter, editors, *Handbook of Modal Logic*, pages 821–868. Elsevier, Amsterdam, 2007.

Kenneth J. Arrow. Rational Choice Functions and Orderings. *Economica*, 26(102): 121–127, 1959.

Sergei Artemov. The Logic of Justification. *The Review of Symbolic Logic*, 1(4): 477–513, 2008.

Robert Audi. *Belief, Justification, and Knowledge*. Wadsworth Publishing, Belmont, 1988.

J. L. Austin. Other Minds. *Proceedings of the Aristotelian Society*, 20:148–187, 1946.

Roberta Ballarin. Validity and Necessity. *Journal of Philosophical Logic*, 34:275–303, 2005.

Alexandru Baltag and Sonja Smets. Probabilistic dynamic belief revision. *Synthese*, 165:179–202, 2008.

Jon Barwise and John Perry. *Situations and Attitudes*. MIT Press, Cambridge, Mass., 1983.

Peter Baumann. Contextualism and the Factivity Problem. *Philosophy and Phenomenological Research*, 76:580–602, 2008.

Kelly Becker. Is Counterfactual Reliabilism Compatible with Higher-Level Knowledge? *Dialectica*, 60(1):79–84, 2006.

Kelly Becker. *Epistemology Modalized*. Routledge, New York, 2007.

Johan van Benthem. *Logical Dynamics of Information and Interaction*. Cambridge University Press, New York, 2011.

Johan van Benthem. Correspondence Theory. In Dov M. Gabbary and Franz Guenthner, editors, *Handbook of Philosophical Logic*, volume 3, pages 325–408. Kluwer, Dordrecht, 2nd edition, 2001.

Johan van Benthem. *Modal Logic for Open Minds*. CSLI Publications, Stanford, 2010.

Johan van Benthem and Fenrong Liu. Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logics*, 17(2):157–182, 2007.

Johan van Benthem and Fernando Raymundo Velázquez-Quesada. The Dynamics of Awareness. *Synthese*, 177:5–27, 2010.

Johan van Benthem, Patrick Girard, and Olivier Roy. Everything Else Being Equal: A Modal Logic for *Ceteris Paribus* Preferences. *Journal of Philosophical Logic*, 38: 83–125, 2009.

Tim Black. Modal and Anti-Luck Epistemology. In Sven Bernecker and Duncan Pritchard, editors, *The Routledge Companion to Epistemology*, pages 187–198. Routledge, New York, 2010.

Michael Blome-Tillmann. Knowledge and Presuppositions. *Mind*, 118(470):241–294, 2009.

Oliver Board. Dynamic Interactive Epistemology. *Games and Economic Behavior*, 49:49–80, 2004.

Laurence BonJour. Nozick, Externalism, and Skepticism. In Steven Luper-Foy, editor, *The Possibility of Knowledge: Nozick and His Critics*, pages 297–313. Rowman & Littlefield, Totowa, 1987.

Georges Bordes. Consistency, Rationality and Collective Choice. *The Review of Economic Studies*, 43(3):451–457, 1976.

Elke Brendel. Why contextualists cannot know they are right: self-refuting implications of contextualism. *Acta Analytica*, 20(2):38–55, 2005.

Jessica Brown. Knowledge and Assertion. *Philosophy and Phenomenological Research*, 81(3):549–566, 2010.

Anthony Brueckner. Klein on Closure and Skepticism. *Philosophical Studies*, 98: 139–151, 1998.

Anthony Brueckner. Strategies for refuting closure for knowledge. *Analysis*, 64(4): 333–335, 2004.

John P. Burgess. Which Modal Models are the Right Ones (for Logical Necessity)? *Theoria*, 18:145–158, 2003.

John P. Burgess. *Philosophical Logic*. Princeton University Press, Princeton, 2009.

Walter Carnielli, Marcelo E. Coniglio, and Joao Marcos. Logics of Formal Inconsistency. In Dov M. Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic*, volume 14, pages 1–93. Springer, Dordrecht, 2007.

David J. Chalmers. The Nature of Epistemic Space. In Andy Egan and Brian Weatherson, editors, *Epistemic Modality*, pages 60–107. Oxford University Press, New York, 2011.

Brian F. Chellas. *Modal Logic: An Introduction.* Cambridge University Press, New York, 1980.

Herman Chernoff. Rational Selection of Decision Functions. *Econometrica*, 22(4): 422–443, 1954.

Alonzo Church. *Introduction to Mathematical Logic.* Princeton University Press, Princeton, 1995/1956.

Stewart Cohen. How to be a Fallibilist. *Philosophical Perspectives*, 2:91–123, 1988.

Stewart Cohen. Contextualist Solutions to Epistemological Problems: Scepticism, Gettier, and the Lotttery. *Australasian Journal of Philosophy*, 76(2):289–306, 1998.

Stewart Cohen. Contextualism, Skepticism, and the Structure of Reasons. *Noûs*, 33: 57–89, 1999.

Stewart Cohen. Contextualism and Skepticism. *Philosophical Issues*, 10:94–107, 2000.

Stewart Cohen. Basic Knowledge and the Problem of Easy Knowledge. *Philosophy and Phenomenological Research*, 65(2):309–329, 2002.

Juan Comesaña. Knowledge and Subjunctive Conditionals. *Philosophy Compass*, 2: 781–791, 2007.

Charles B. Cross. Antecedent-Relative Comparative World Similarity. *Journal of Philosophical Logic*, 37:101–120, 2008.

Keith DeRose. Solving the Skeptical Problem. *The Philosophical Review*, 104(1): 1–52, 1995.

Keith DeRose. How Can We Know that We're Not Brains in Vats? *The Southern Journal of Philosophy*, 38:121–148, 2000.

Keith DeRose. Sosa, Safety, Sensitivity, and Skeptical Hypotheses. In John Greco, editor, *Ernest Sosa and His Critics*, pages 22–41. Blackwell, Malden, 2004.

Keith DeRose. *The Case for Contextualism*. Oxford University Press, New York, 2009.

Keith DeRose. Contextualism, Contrastivism, and X-Phi Surveys. *Philosophical Studies*, 156:81–110, 2011.

Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi. *Dynamic Epistemic Logic*. Springer, Dordrecht, 2008.

Fred Dretske. Epistemic Operators. *The Journal of Philosophy*, 67(24):1007–1023, 1970.

Fred Dretske. Conclusive Reasons. *Australasian Journal of Philosophy*, 49(1):1–22, 1971.

Fred Dretske. The Pragmatic Dimension of Knowledge. *Philosophical Studies*, 40: 363–378, 1981.

Fred Dretske. Externalism and Modest Contextualism. *Erkenntnis*, 61:173–186, 2004.

Fred Dretske. The Case against Closure. In Matthias Steup and Ernest Sosa, editors, *Contemporary Debates in Epistemology*, pages 13–25. Blackwell, Malden, 2005.

J. Michael Dunn. Intuitive Semantics for First-Degree Entailments and 'Coupled Trees'. *Philosophical Studies*, 29(3):149–168, 1976.

Kfir Eliaz and Efe A. Ok. Indifference or indecisiveness? Choice-theoretic foundations of incomplete preferences. *Games and Economic Behavior*, 56:61–86, 2006.

Gareth Evans. Reference and Contingency. *The Monist*, 62(2):161–189, 1979.

Frederic B. Fitch. A Logical Analysis of Some Value Concepts. *The Journal of Symbolic Logic*, 28(2):135–142, 1963.

Melvin Fitting. Reasoning with Justifications. In David Makinson, Jacek Malinowski, and Heinrich Wansing, editors, *Towards Mathematical Philosophy*, volume 28 of *Trends in Logic*, pages 107–123. Springer, Dordrecht, 2009.

Nir Friedman and Joseph Y. Halpern. On the Complexity of Conditional Logics. In Jon Doyle, Erik Sandwell, and Pietro Torasso, editors, *Proceedings of the Fourth International Conference on Principles of Knowledge Representation and Reasoning (KR'94)*, pages 202–213. Morgan Kaufman, San Francisco, 1994.

Alan H. Goldman. A Note on the Conjunctivity of Knowledge. *Analysis*, 36(1):5–9, 1975.

Alvin I. Goldman. Discrimination and Perceptual Knowledge. *The Journal of Philosophy*, 73(20):771–791, 1976.

Alvin I. Goldman. *Epistemology and Cognition*. Harvard University Press, Cambridge, Mass., 1986.

Paul Grice. *Studies in the Way of Words*. Harvard University Press, Cambridge, Mass., 1989.

Joseph Y. Halpern. The effect of bounding the number of primitive propositions and the depth of nesting on the complexity of modal logic. *Artificial Intelligence*, 75 (2):361–372, 1995.

Joseph Y. Halpern. Set-theoretic completeness for epistemic and conditional logic. *Annals of Mathematics and Artificial Intelligence*, 26:1–27, 1999.

Joseph Y. Halpern and Riccardo Pucella. Dealing with logical omniscience: Expressiveness and pragmatics. *Artificial Intelligence*, 175:220–235, 2011.

Joseph Y. Halpern and Leandro Chaves Rêgo. Characterizing the NP-PSPACE Gap in the Satisfiability Problem for Modal Logic. *Journal of Logic and Computation*, 17(4):795–806, 2007.

Gilbert Harman and Brett Sherman. Knowledge, Assumptions, Lotteries. *Philosophical Issues*, 14(1):492–500, 2004.

John Hawthorne. *Knowledge and Lotteries*. Oxford University Press, New York, 2004a.

John Hawthorne. Replies. *Philosophical Issues*, 14(1):510–523, 2004b.

John Hawthorne. The Case for Closure. In Matthias Steup and Ernest Sosa, editors, *Contemporary Debates in Epistemology*, pages 26–43. Blackwell, Malden, 2005.

Mark Heller. Relevant Alternatives. *Philosophical Studies*, 55:23–40, 1989.

Mark Heller. Relevant Alternatives and Closure. *Australasian Journal of Philosophy*, 77(2):196–208, 1999a.

Mark Heller. The Proper Role for Contextualism in an Anti-Luck Epistemology. *Noûs*, 33:115–129, 1999b.

Herbert E. Hendry and M.L. Pokriefka. Carnapian Extensions of S5. *Journal of Philosophical Logic*, 14:111–128, 1985.

Jaakko Hintikka. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca, 1962.

Wesley H. Holliday. Epistemic Logic, Relevant Alternatives, and the Dynamics of Context. In Daniel Lassiter and Maria Slavkovic, editors, *New Directions in Logic, Language, and Computation*, volume 7415 of *Lecture Notes in Computer Science*, pages 109–129. Springer, Dordrecht, 2012.

Wesley H. Holliday. Epistemic Closure and Epistemic Logic I: Relevant Alternatives and Subjunctivism. *Journal of Philosophical Logic*, 2013a. Forthcoming.

Wesley H. Holliday. Fallibilism and Multiple Paths to Knowledge. In Tamar Szabó Gendler and John Hawthorne, editors, *Oxford Studies in Epistemology*, volume 5. Oxford University Press, New York, 2013b. Forthcoming.

Wesley H. Holliday. Epistemic Closure and Epistemic Logic II: A New Framework for Fallibilism. Manuscript, 2013c.

Wesley H. Holliday and Thomas F. Icard, III. Moorean Phenomena in Epistemic Logic. In Lev Beklemishev, Valentin Goranko, and Valentin Shehtman, editors, *Advances in Modal Logic*, volume 8, pages 178–199. College Publications, London, 2010.

Wesley H. Holliday and John Perry. Roles, Rigidity, and Quantification in Epistemic Logic. In Alexandru Baltag and Sonja Smets, editors, *Johan F. A. K. van Benthem on Logical and Informational Dynamics*. Springer, Dordrecht, 2013. Forthcoming.

Wesley H. Holliday, Tomohiro Hoshi, and Thomas F. Icard, III. Schematic Validity in Dynamic Epistemic Logic: Decidability. In Hans van Ditmarsch, Jérôme Lang, and Shier Ju, editors, *Proceedings of the Third International Workshop on Logic, Rationality and Interaction (LORI-III)*, volume 6953 of *Lecture Notes in Artificial Intelligence*, pages 87–96. Springer, Dordrecht, 2011.

Wesley H. Holliday, Tomohiro Hoshi, and Thomas F. Icard, III. A Uniform Logic of Information Dynamics. In Thomas Bolander, Torben Braüner, Silvio Ghilardi, and Lawrence Moss, editors, *Advances in Modal Logic*, volume 9, pages 348–367. College Publications, London, 2012.

Jonathan Ichikawa. Quantifiers and epistemic contextualism. *Philosophical Studies*, 155:383–398, 2011.

David Kaplan. A Problem in Possible World Semantics. In Walter Sinnott-Armstrong, Diana Raffman, and Nicholas Asher, editors, *Modality, morality, and belief: essays in honor of Ruth Barcan Marcus*, pages 41–52. Cambridge University Press, Cambridge, 1995.

Jeffrey C. King. What in the world are ways things might have been? *Philosophical Studies*, 133:443–453, 2007.

Peter Klein. Skepticism and Closure: Why the Evil Genius Argument Fails. *Philosophical Topics*, 23(1):213–236, 1995.

Saul Kripke. Semantical Considerations on Modal Logic. *Acta Philosophica Fennica*, 16:83–94, 1963.

Saul Kripke. Nozick on Knowledge. In *Philosophical Troubles: Collected Papers*, volume 1, pages 162–224. Oxford University Press, New York, 2011.

Jonathan L. Kvanvig. Closure Principles. *Philosophy Compass*, 1(3):256–267, 2006.

Henry E. Kyburg, Jr. *Probability and the Logic of Rational Belief*. Wesleyan University Press, Middletown, 1961.

Maria Lasonen-Aarnio. Single Premise Deduction and Risk. *Philosophical Studies*, 141:157–173, 2008.

Krista Lawlor. Living without Closure. *Grazer Philosophische Studien*, 69:25–49, 2005.

Krista Lawlor. *Assurance: An Austinian View of Knowledge and Knowledge Claims*. Oxford University Press, New York, 2013.

Hector J. Levesque. A Logic of Implicit and Explicit Belief. *Proceedings of AAAI-84*, pages 198–202, 1984.

David Lewis. Completeness and Decidability of Three Logics of Counterfactual Conditionals. *Theoria*, pages 74–85, 1971.

David Lewis. *Counterfactuals*. Basil Blackwell, Oxford, 1973.

David Lewis. Possible-World Semantics for Counterfactual Logics: A Rejoinder. *Journal of Philosophical Logic*, 6:359–363, 1977.

David Lewis. Scorekeeping in a Language Game. *Journal of Philosophical Logic*, 8: 339–359, 1979.

David Lewis. Ordering Semantics and Premise Semantics for Counterfactuals. *Journal of Philosophical Logic*, 10:217–234, 1981.

David Lewis. *On the Plurality of Worlds*. Basil Blackwell, Oxford, 1986.

David Lewis. Statements Partly About Observation. *Philosophical Papers*, 17:1–31, 1988.

David Lewis. Elusive Knowledge. *Australasian Journal of Philosophy*, 74(4):549–567, 1996.

Barry M. Loewer. Cotenability and Counterfactual Logics. *Journal of Philosophical Logic*, 8(1):99–115, 1979.

Steven Luper-Foy. The Epistemic Predicament: Knowledge, Nozickian Tracking, and Scepticism. *Australasian Journal of Philosophy*, 62(1):26–49, 1984.

Steven Luper-Foy. Introduction. In Steven Luper-Foy, editor, *The Possibility of Knowledge: Nozick and His Critics*, pages 1–16. Rowman & Littlefield, Totowa, 1987.

John MacFarlane. The Assessment-Sensitivity of Knowledge Attributions. *Oxford Studies in Epistemology*, 1:197–233, 2005.

D. C. Makinson. The Paradox of the Preface. *Analysis*, 25:205–207, 1965.

John C. Mayer. A Misplaced Thesis of Conditional Logic. *Journal of Philosophical Logic*, 10(2):235–238, 1981.

Colin McGinn. The Concept of Knowledge. *Midwest Studies in Philosophy*, 9:529–554, 1984.

C. A. Middelburg. A Survey of Paraconsistent Logics. 2011. URL `http://arxiv.org/abs/1103.4324v1`.

Peter Murphy. Closure Failures for Safety. *Philosophia*, 33:331–334, 2005.

Peter Murphy. A Strategy for Assessing Closure. *Erkenntnis*, 65:365–383, 2006.

Robert Nozick. *Philosophical Explanations*. Harvard University Press, Cambridge, Mass., 1981.

Donald Nute. Counterfactuals and the Similarity of Words [sic]. *The Journal of Philosophy*, 72(21):773–778, 1975.

Donald Nute. Simplification and Substitution of Counterfactual Antecedents. *Philosophia*, 7:317–325, 1978.

Donald Nute. Conversational Scorekeeping and Conditionals. *Journal of Philosophical Logic*, 9:153–156, 1980.

Mike Oaksford and Keith Stenning. Reasoning With Conditionals Containing Negated Constituents. *Journal of Experimental Psychology*, 18(4):835–854, 1992.

Christos H. Papadimitriou. *Computational Complexity*. Addison-Wesley, Reading, 1994.

William T. Parry. Ein Axiomensystem für eine neue Art von Implikation (Analytische Implikation). *Ergebnisse eines mathematischen Kolloquiums*, 4:5–6, 1933.

William T. Parry. Analytic Implication: Its History, Justification and Varieties. In John Norman and Richard Sylvan, editors, *Directions in Relevant Logic*, volume 1 of *Reason and Argument*, pages 101–118. Kluwer, Dordrecht, 1989.

John Perry. Possible Worlds and Subject Matter. 1989. Reprinted in Perry 2000, 145-160.

John Perry. Evading the Slingshot. 1996. Reprinted in Perry 2000, 287-301.

John Perry. *The Problem of the Essential Indexical and Other Essays*. CSLI Publications, Stanford, 2000.

Graham Priest. Paraconsistent Set Theory. In David DeVidi, Michael Hallett, and Peter Clark, editors, *Logic, Mathematics, Philosophy: Vintage Enthusiasms*, volume 75 of *The Western Ontario Series in Philosophy of Science*, pages 153–169. Springer, Dordrecht, 2011.

Duncan Pritchard. *Epistemic Luck*. Oxford University Press, New York, 2005.

Duncan Pritchard. Sensitivity, Safety, and Anti-Luck Epistemology. In John Greco, editor, *The Oxford Handbook of Skepticism*, pages 437–455. Oxford University Press, New York, 2008.

James Pryor. Highlights of Recent Epistemology. *British Journal of the Philosophy of Science*, 52:95–124, 2001.

Hans Rott. *Change, Choice, and Inference: A Study of Belief Revision and Non-monotonic Reasoning*. Oxford University Press, New York, 2001.

Sherrilyn Roush. *Tracking Truth: Knowledge, Evidence, and Science*. Oxford University Press, New York, 2005.

Sherrilyn Roush. Closure on Skepticism. *The Journal of Philosophy*, 107(5):243–256, 2010.

Sherrilyn Roush. Sensitivity and Closure. In Kelly Becker and Tim Black, editors, *The Sensitivity Principle in Epistemology*, pages 242–268. Cambridge University Press, New York, 2012.

Patrick Rysiew. The Context-Sensitivity of Knowledge Attributions. *Noûs*, 35(4): 477–514, 2001.

Patrick Rysiew. Motivating the Relevant Alternatives Approach. *Canadian Journal of Philosophy*, 36(2):259–279, 2006.

Nathan Salmon. The Logic of What Might Have Been. *The Philosophical Review*, 98 (1):3–34, 1989.

Amartya K. Sen. Choice Functions and Revealed Preference. *The Review of Economic Studies*, 38(3):307–317, 1971.

Brett Sherman and Gilbert Harman. Knowledge and Assumptions. *Philosophical Studies*, 156(1):131–140, 2011.

Brian Skyrms. The Explication of "X knows that p". *The Journal of Philosophy*, 64 (12):373–389, 1967.

Ernest Sosa. Postscript to "Proper Functionalism and Virtue Epistemology". In Jonathan Kvanvig, editor, *Warrant in Contemporary Epistemology*, pages 271–281. Rowman & Littlefield, Totowa, 1996.

Ernest Sosa. How to Defeat Opposition to Moore. *Noûs*, 33(13):141–153, 1999.

Ernest Sosa. Relevant Alternatives, Contextualism Included. *Philosophical Studies*, 119:35–65, 2004.

Robert Stalnaker. A Theory of Conditionals. In Nicholas Rescher, editor, *Studies in Logical Theory*, volume 2, pages 98–112. Basil Blackwell, Oxford, 1968.

Robert Stalnaker. The Problem of Logical Omniscience I. *Synthese*, 89:425–440, 1991.

Robert Stalnaker. Common Ground. *Linguistics and Philosophy*, 25:701Ð721, 2002.

Robert C. Stalnaker. *Inquiry*. MIT Press, Cambridge, Mass., 1984.

Jason Stanley. Fallibilism and concessive knowledge attributions. *Analysis*, 65(2): 126–131, 2005.

G.C. Stine. Skepticism, Relevant Alternatives, and Deductive Closure. *Philosophical Studies*, 29:249–261, 1976.

Barry Stroud. *The Significance of Philosophical Scepticism*. Oxford University Press, New York, 1984.

John Turri. Epistemic Invariantism and Speech Act Contextualism. *Philosophical Review*, 119(1):77–95, 2010.

Moshe Y. Vardi. On the Complexity of Epistemic Reasoning. In *Proceedings of the Fourth Annual Symposium on Logic in Computer Science*, pages 243–252. IEEE Publishing, 1989.

Fernando Raymundo Velázquez-Quesada. Inference and Update. *Synthese*, 169:283–300, 2009.

Fernando Raymundo Velázquez-Quesada. *Small Steps in the Dynamics of Information*. PhD thesis, University of Amsterdam, 2011. ILLC Dissertation Series DS-2011-02.

Jonathan Vogel. Tracking, Closure, and Inductive Knowledge. In Steven Luper-Foy, editor, *The Possibility of Knowledge: Nozick and His Critics*, pages 197–215. Rowman & Littlefield, Totowa, 1987.

Jonathan Vogel. The New Relevant Alternatives Theory. *Noûs*, 33:155–180, 1999.

Jonathan Vogel. Reliabilism Leveled. *The Journal of Philosophy*, 97(11):602–623, 2000.

Jonathan Vogel. Subjunctivitis. *Philosophical Studies*, 134:73–88, 2007.

Ted A. Warfield. When epistemic closure does and does not fail: a lesson from the history of epistemology. *Analysis*, 64(1):35–41, 2004.

Matt Weiner. Norms of Assertion. *Philosophy Compass*, 2(2):187–195, 2007.

Timothy Williamson. *Knowledge and its Limits*. Oxford University Press, New York, 2000.

Timothy Williamson. Comments on Michael Williams' "Contextualism, Externalism, and Epistemic Standards". *Philosophical Studies*, 103:25–33, 2001.

Timothy Williamson. Probability and Danger. *The Amherst Lecture in Philosophy*, 4:1–35, 2009. URL `http://www.amherstlecture.org/williamson2009/`.

Timothy Williamson. Interview. In Vincent Hendricks and Olivier Roy, editors, *Epistemic Logic: 5 Questions*, pages 249–261. Automatic Press, Copenhagen, 2010.

Crispin Wright. Contextualism and Scepticism: Even-Handedness, Factivity, and Surreptitiously Raising Standards. *The Philosophical Quarterly*, 55(219):236–262, 2005.

Crispin Wright. On Epistemic Entitlement (II): Welfare State Epistemology. In Dylan Dodd and Elia Zardini, editors, *Contemporary Perspectives on Scepticism and Perceptual Justification*. Oxford University Press, New York, forthcoming.

Stephen Yablo. Knowing About Things. 2012a. URL `http://www.mit.edu/~yablo/home/Papers.html`.

Stephen Yablo. Aboutness Theory. 2012b. URL `http://www.mit.edu/~yablo/home/Papers.html`.