MODELS OF KNOWLEDGE FOR RESOURCE BOUNDED AGENTS

by

Jacob N. Caton

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF PHILOSOPHY

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2012

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Jacob N. Caton
entitled Models of Knowledge for Resource Bounded Agents
and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Date: 28 March 2012
Stewart Cohen

Date: 28 March 2012
Terry Horgan

Date: 28 March 2012
Juan Comesaña

Date: 28 March 2012

Date: 28 March 2012

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.
I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Date: 28 March 2012
Dissertation Director: Stewart Cohen

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Jacob N. Caton

ACKNOWLEDGEMENTS

I would like to express my sincerest thanks and appreciation to my committee members: Stew Cohen, Terry Horgan, and Juan Comesaña. I am grateful for the many discussions, meetings, and lunches we had over the years, for the time they gave providing feedback, and for the support and encouragement they gave at every stage of the dissertation. And in their own individual way, for providing me with excellent models of what it is to be a philosopher: Stew for his tireless demand for clarity and precision, Terry for his broad and patient view of philosophical inquiry (and his appreciation of philosophical puzzles), and Juan for his ability to see and remain focused on what is of central importance in a philosophical claim.

For their valuable discussions and comments at various stages of this work, I would like to thank Adam Arico, Nathan Ballantyne, Anne Baril, Ian Evans, Don Fallis, David Glick, Rachana Kamtekar, Keith Lehrer, Stephen Lenhart, Theresa Lopez, Shuan Nichols, Joseph Tolliver, and Orlin Vakarelov.

I would also like to acknowledge a deep debt of gratitude to John Pollock. John was a friend and my first philosophical mentor. His intelligence, philosophical depth, humility, and joyful outlook on life continue to be a source of inspiration. I miss discussing philosophy with him. I began this dissertation with John as my advisor and I hope he would have found some of my results interesting.

Thanks to my parents, Jerald and Roberta Caton, for their support, encouragement, and guidance.

And importantly, to Tiffany, for always being there for *me*. In the language of this dissertation, she makes me aware of the possibilities.

DEDICATION

To my parents, Jerald and Roberta Caton

TABLE OF CONTENTS

TABLE OF CONTENTS – *Continued*

TABLE OF CONTENTS – *Continued*

ABSTRACT

We know things about the world in spite of our cognitive limitations and imperfections. Occasions of stress impact memory retrieval, resources for attention can be depleted by non-epistemic factors, and our visual system has limited resolution and discriminatory ability. Yet we know many propositions, ranging from the mundane to the arcane, and we often are able to know that we know these things. In this dissertation I explore the relationship between our cognitive limitations and the limits to what we know, and what we know that we know.

I begin by considering a simple model of knowledge. Because it is difficult (perhaps impossible) to have intuitions about many higher-order or iterative knowledge claims ("I know that you know that she knows that I know that ..."), a modeling approach can help clarify and explain how various cognitive limitations impact knowledge and higher-order knowledge. In Chapter 2 I discuss the epistemic requirements for the rational coordination of our actions. While it may seem that coordination is rational only if each coordinating member has what may be called "common knowledge" of some relevant proposition, the model of knowledge I employ helps show the informational complexity of common knowledge. I argue that common knowledge is unattainable. In Chapter 3 I discuss epistemic closure. Perfectly ideal agents may know every deductive consequence of what they know, but if the aim is to understand how deduction extends human knowledge then it is necessary to model our cognitive access to information. In Chapter 4 I turn to to the issue of higher-order or iterative knowledge. I argue that memory limitations and various information processing errors all result in failures of higher-order knowledge. The argument I give does not require epistemic closure or a principle of self-knowledge. I conclude, in Chapter 5, by discussing interpretive issues for models of knowledge and I discuss our awareness of what we know and what we do not know.

CHAPTER 1

A MODEL FOR KNOWLEDGE

1.1 Introduction

We know many things. I know that it is sunny outside now, and I also know that it is usually sunny in Tucson. I know that it has been sunny here in the past, and I know that it will be sunny tomorrow.

For some, these claims are unremarkable. Without focused philosophical attention, it may seem commonplace that we know these things. Yet it is also commonplace to note that we are not perfect reasoners—we have limited memories, our eyesight is imperfect, we have limited attentional resources—to name only a few imperfections. Somehow, it seems that we manage to know what we know in spite of our cognitive limitations. As an epistemologist, I want to know how far our limited cognitive resources can work to produce knowledge (what is the extent of our knowledge) and why this is so.

The rub, however, is that an investigation of human knowledge requires utilization of the same imperfect cognitive resources that are responsible for the deficits in our knowledge. Such a difficult situation is particularly acute when we consider what we do not know. There are some things that we do not know because we fail to pay appropriate attention. But then we often do not know that we do not know. Our failures of attention rob us from knowing, but they also seem to rob us from knowing that we do not know. In this way, the boundary separating what we know from what we do not know appears difficult to chart.

A standard approach to answering questions about the limits of human knowledge proceeds by way of philosophical intuition, but philosophical intuition is also

potentially impacted by our limited cognitive resources. It is difficult (perhaps impossible) to have an intuition about an infinite series of knowledge claims and it is difficult (perhaps impossible) to have an intuition about knowledge claims framed in grammatically complicated ways. As it turns out, questions about common knowledge and iterated knowledge often have these features. As such, philosophical intuition should be fortified by other means. In this chapter I will introduce a model for knowledge that will help reveal the relationship between what we know and our cognitive limitations.

1.2 Models, Definitions, and Semantic Competence

I will begin by introducing a simple model for knowledge, one that will be improved upon throughout subsequent chapters. The core notion of this model can be interpreted in several ways: to know a proposition p is to eliminate all non-p states; or, knowledge is truth in all possible worlds. Though there are substantive differences, these interpretations suggest an important relationship between possibility and knowledge: when one knows that p, it is not possible that p is false; when it is possible that p, one does not know that p is false.

I want to stress the difference between a *model* of knowledge and an *analysis* of knowledge. My project is not one of analysis. I do not aim to give necessary and sufficient conditions for knowledge, nor do I aim to decompose the concept of knowledge into simpler parts. To model a phenomenon is to give an idealized representation with an aim to reveal its structure *in a particular dimension*, but remain silent about other dimensions. As a motivating example, consider the familiar device of a town model on display at many museums. A model of a town is a small-scale figure that reveals various spatial relations.[1] One may learn from a town model whether the bank is near the library. Or one may learn from a town model whether the school is north of the police station. But, from the model, one will probably *not*

[1]Such models are often called "material models" to highlight that they are physical objects.

learn whether the bank building is heavier than the library. This is because most town models aim only to capture various geographical or spatial properties.

Next, consider several models from the sciences. A simple physical model of forces might, with a few assumptions, show how force is related to properties of mass and acceleration. Such introductory models might be unrealistic as a complete theory of the nature of forces, but, at the appropriate resolution, these models help reveal general features of the relationship between force and mass. So too for the social sciences. A model of GDP (gross domestic product) might show how savings and the interest rate affect the output of a country. Again, such an introductory model might not serve as a complete story of GDP. Yet progress is made when it can be shown that there exists a general relationship between interest rates and output. Further progress may be achieved when the introductory model is improved.

1.2.1 Epistemic Models

Epistemologists do not typically work with models. I want to suggest two ways of thinking about epistemic models and argue that they can help address important epistemic questions.

Epistemic models are not like material models of bridges or automobiles. One understanding of an epistemic model is as an "idealized model". As Frigg and Hartmann (2012) describe them, idealized models are deliberate simplifications of a complicated phenomenon, with an aim to increase the tractability of a problem. What, then, might be idealized in an epistemic model? Some epistemologists have sought to idealize the human knower as an "ideally rational agent". Everyday human knowers are complicated entities, with complex psychologies and cognitive limitations. Just as frictionless planes are easier to understand than structures found in the real world, ideal agents are (comparably) easier to understand than actual subjects found in the real world. For purposes of simplification, it is assumed that ideal agents do not make mistakes in reasoning, have perfect memo-

ries, and do not feature information processing limitations. If ideal knowers could be captured through the use of a model, then realistic human knowers could be approximated as simplifying assumptions (e.g., no memory failures) are removed. Such a process has the potential to help reveal the relationship between cognitive limitations and their corresponding epistemic consequences.

Epistemic models might also be understood as providing kinds of analogies for use in broad analogical reasoning. Just as some have analogized the mind as a kind of computer, human knowers might be thought to be similar to information processors. Often, work done by theoretical computer scientists in epistemic logic seems motivated by such an analogy.[2] To take a concrete case, if one is interested in the question of how knowledge is collected and shared by groups of agents, one might note the similarities between groups of human agents and groups of distributed computer processing systems. Each may be judged as facing a principally informational task. But because distributed processors are well-studied, results in theoretical computer science may be applied (by way of analogical inference) to questions about the human knower.

However, in either above interpretation, epistemic models are able to operate at a remove from the messiness of traditional epistemic theorizing. Intuitions about cases are often frail, hard to interpret, or unclear in their value. Epistemic principles can be vague or, when repeatedly amended and fortified, hard to understand (and read). Epistemic models offer an alternative approach. The kinds of epistemic model I'm concerned with are formal models, so many of the interrelations between concepts and corresponding results can be expressed and captured formally. Such formalisms provide an exacting yet expressive language for investigation. Epistemic models may sacrifice perfect reality for simplicity, but the gain is precision.

[2]See Halpern and Moses (1990) for one such example. Papers published by the research group of Ronald Fagin, Joseph Halpern, Yoram Moses, and Moshe Vardi often use this analogy.

1.2.2 Definitions and Semantic Competence

Models and definitions are not, necessarily, incompatible pursuits. Yet, they typically offer promise of different insight. A significant part of contemporary epistemology has been concerned with the question "what is knowledge". A model of knowledge won't, on its own, answer this question (though it may help inform our understanding of knowledge). Likewise, a model of GDP won't answer the question "what is the output of an economy" (a question addressed by some economists). Yet many economists work with models of GDP, without a unanimously shared definition of economic output, because models can show the structural relationship (or interrelation) between properties such as saving rates and interest rates.

Similarly, there are structural questions to ask of knowledge (how various other concepts or properties are related to knowledge). We may wish to inquire how knowledge and action are related (it is plausible to think that not knowing some relevant proposition makes it practically irrational to act on the proposition). We may wish to inquire about how memory or unawareness affects knowledge and our knowledge of our knowledge. We may wish to inquire about how deductive inference can extend knowledge. Models can shed light on these questions in the absence of a universally agreed upon definition of knowledge. It is an aim of this dissertation to show how this is possible.

Yet, one may wonder how we could ever answer any of these above questions without a definition of knowledge. How can a project to model knowledge ever be successful if we don't know what knowledge is? My response is that we know many things about knowledge, even though it is difficult to articulate just what it is that we know. A successful model of knowledge should begin with a good approximation, captured by some of the many things we do know about knowledge, and build toward better and better approximations.

So, what do we know about knowledge? Behavioral psychologists have observed that the word 'knows' and its cognates (know, knew, known) are the most common

words American children use to describe the mental states of themselves or others.[3] Children use 'knows' more than 'thinks' or 'believes'. So, from our beginnings, we develop a semantic competence with the word 'knows'—we know how to use the term. And this competence absolutely extends into adult life. Appraisals of knowledge are common features of life. Juries are often charged with the task of determining whether a defendant had an intention or knowledge of wrongdoing (*mens rea*). Journalists and political commissions attempt to determine what a politician knew, and when she knew it. A supervisor might ask, "do you know where the Penske file is"? A waiter might be asked if he knows when the soup will be ready. In each of these pursuits, it would be aberrant for someone to respond, "what do you mean by 'know' "? We know how to correctly use the term 'know'—we have semantic competence for the term (again, even though we might not be able to articulate a theory of the underpinnings of this competence).

1.2.3 Two Principles

Given our semantic competence with 'knows', consider a case. Suppose Adam and Brett are in the dining room when Adam asks Brett if she knows where his keys are (he is worried he misplaced them). Brett says she knows that Adam left them in the laundry room. But Adam replies that he remembers picking them up off the dryer, so they can't be there. Brett then retracts her earlier claim when she says: "I guess I don't know where your keys are".

Such a case is highly plausible and it helps support an initial view about the relationship between knowledge and possibility. Brett's retraction could be modeled by the following principle:

(1) If p is not possible for S, then S does not know that p.

[3]See Bartsch and Wellman (1995).

When Brett learns that the keys couldn't be in the laundry room, she admits that she doesn't know that they are.

However, Brett and Adam's conversation might have followed a different path. Suppose, instead, that Adam asks Brett if she knows where his keys are. Brett responds: "they might be in the laundry room, but they also might still be in your car—I guess I don't know where they are". Again, such a case is highly plausible and representative of our everyday use of 'knows'. Brett's response in this second case could be modeled by this principle:

(2) If an alternative q (to p) is possible for S, then S doesn't know p.

When Brett remarks that the keys might be in the laundry room, we take her to acknowledge that the keys cannot both be in the laundry room and the car at the same time. She admits that the keys might be in the laundry room, so she doesn't know that they are in his car.[4]

I don't take principles (1) and (2) as definitive, but they seem reasonable enough starting places. Principles (1) and (2) may or may not be jointly sufficient or necessary for a definition of knowledge. But, for purposes of building a model, these principles serve as good first approximations. And these principles support the core notion of knowledge that I gave in the opening paragraph of section 1.2: to know a proposition p is to eliminate all non-p states. I next turn to a set-theoretic model of knowledge that builds on principles (1) and (2), and aims to capture this core notion of knowledge.

[4]She actually says something stronger—she doesn't know where they are. This entails that she doesn't know that the keys are in his car.

1.3 A Set-Theoretic Model of Knowledge

Begin with a set Ω of states.[5] Take a state ω to be a "full description of the world", one that resolves all matters of fact relevant to the situation. States are mutually exclusive such that no two states may obtain, and one state is the "actual state". Next, define a function P on Ω that returns, for each $\omega \in \Omega$, a non-empty subset of states $P(\omega)$. The interpretation of P is epistemic: the set $P(\omega)$ includes all states that are epistemically possible for the agent at ω. So, if $\omega' \in P(\omega)$, then ω' is possible for the agent at ω.

Consider an example. Suppose an agent S considers the outcome of rolling a six-sided die. Suppose the die is rolled beyond S's sight in an adjacent room, and that the die lands on 2. Initially, we may suppose that there are only six possible states of the world. That is, $\Omega = \{1, 2, 3, 4, 5, 6\}$. As we assumed, the actual state of world is now 2, because the die landed on 2. Before S learned of the outcome of the roll, it seems natural to assume that every state was possible for S, or $P(2) = \{1, 2, 3, 4, 5, 6\}$. Next, suppose that S is told that the die landed on 2. If S updates correctly, then S's possibility function changes to $P(2) = \{2\}$. That is, when S is told that the die landed on 2 she no longer thinks that it is possible that the die landed on 3, and so forth. Suppose, instead, that S makes a mistake: when S is told that the die landed on 2 she mistakenly hears "the die landed on 6". In this case, S's possibility function is $P(2) = \{6\}$.

Next, define an "event". Any set of states E is called "the event E". Say that an event E obtains if and only if the actual state $\omega \in E$. In many cases, a random set of states will not have much cognitive significance. Suppose, again, that $\Omega = \{1, 2, 3, 4, 5, 6\}$. Let $E = \{1, 2, 3\}$. One may simply view E as the disjunction of states 1, 2, and 3. E obtains if and only if either a 1, a 2, or a 3 is rolled. Or, alternatively, one may interpret E as the event "the number rolled is less than 4".

[5] My presentation and discussion of the set-theoretic model of knowledge is adapted from Rubinstein (1998).

Or, alternatively, one may interpret E as "the second, third, or fourth number in the Fibonacci sequence is rolled".

Call a pair (Ω, P) an "information structure".[6] Informally, an information structure is an epistemic model of an agent. That is, an information structure for S is intended to model what an agent knows. Further, what an agent does or does not know is represented by events.

> **Set-Theoretic Definition of Knowledge**: Let (Ω, P) be an information structure. The event E is known at state ω if and only if $P(\omega) \subseteq E$.

The plausibility of this definition derives from principles (1) and (2) from section 1.2. The major interpretation of this definition is that the condition "$P(\omega) \subseteq E$" obtains for the agent S when and only when the only possible states, for the agent, are ways that are E. Certainly, the set-theoretic definition of knowledge is stronger than principles (1) and (2) (the definition entails the principles, but not conversely). Yet, the set-theoretic definition of knowledge has many plausible features.

Consider, again, the above case. Suppose S learns that the die landed on 2. Initially, S's possibility function was $P(2) = \{1, 2, 3, 4, 5, 6\}$ (before she was told of the result), but, after learning the result, suppose that $P(2) = \{2\}$. In this case, it is intuitive that S knows that the die landed on 2. This accords with the set-theoretic definition of knowledge. The event that the die landed on 2 is $E = \{2\}$. Next, check that S knows E when $P(2) = \{2\}$. By the set-theoretic definition of knowledge, S knows E because $P(2) = \{2\} = E$.

Now, suppose that S mistakenly heard "the die landed on 6" when, in fact, she was told that the die landed on 2. In this case, her possibility function returns $P(2) = \{6\}$. It is easy to show that in this case S doesn't know that the die landed on 2. Again, $E = \{2\}$, and, clearly, $P(2) \nsubseteq E$, because $6 \in P(2)$ but $6 \notin E$.

[6] In game theory and mathematical economics, information structures are sometimes called "Aumann structures", after the economist Robert Aumann.

The core idea behind the set-theoretic definition of knowledge has a distinctive epistemic pedigree: it is the idea that knowledge is modal in character. To know an event E is to, in some sense, eliminate all non-E states as possible.

As I've given the model so far, much is left open to interpretation and refinement. What is it to eliminate a possibility? Are we modeling the agent's epistemic perspective or our perspective of the agent? I address these questions in Chapter 5. But, next, I want to address several properties of the model.

First, I've taken an agent's possibility function as basic to the model, and possibility functions determine what an agent knows. We can therefore uncover the "knowledge operator" K with the set

$$K(E) = \{\omega : P(\omega) \subseteq E\}.$$

The interpretation of $K(E)$ is the event "the agent knows E", which, intuitively, may obtain in any state ω where the agent's possibility function is such that $P(\omega) \subseteq E$.

The knowledge operator K has several important properties. Regardless of any assumptions about P, the following hold for K:

(K0): If $E \subseteq F$, then $K(E) \subseteq K(F)$.

(K0′): $K(E \cap F) = K(E) \cap K(F)$.

(K0″): $K(\Omega) = \Omega$.

Each of the properties (K0)–(K0″) are worth considering in more detail.

Property (K0) is a closure property on Knowledge. Read (K0) as "if event E entails F, then if an agent knows E she knows F". I discuss closure properties, especially (K0) and variants of (K0) in Chapter 3. Yet, it should be clear that (K0) follows from the set-theoretic definition of knowledge, *without* any assumptions on P. Suppose that F obtains whenever E obtains (so, $E \subseteq F$). This is to suppose that E entails F. Suppose, as well, that the event $K(E)$ obtains. By the definition of an

event, $K(E)$ obtains if and only if the actual state $\omega \in K(E)$. By the set-theoretic definition of knowledge, $P(\omega) \subseteq E$. But because $E \subseteq F$, it follows that $P(\omega) \subseteq F$, so $\omega \in K(F)$, by the set-theoretic definition of knowledge.

Next, consider (K0′). This property is read as "knowing E and F is equivalent to knowing E and knowing F". While less controversial than (K0), this property also follows from the set-theoretic definition of knowledge without any assumptions on P. For an equivalence, suppose that S knows E and F. By the set-theoretic definition of knowledge, $\omega \in K(E \cap F)$ if and only if $P(\omega) \subseteq (E \cap F)$ if and only if $P(\omega) \subseteq E$ *and* $P(\omega) \subseteq F$, but this means that $K(E)$ and $K(F)$ obtain, so $\omega \in K(E)$ and $\omega \in K(F)$.

Finally, consider (K0″). One way to interpret (K0″) is that an agent always knows that "something happens" (i.e., some state obtains). This doesn't seem problematic. Yet, another way to interpret (K0″) is that the agent has knowledge of the state space. This seems more troubling. It is plausible to think that an agent knows the possibilities of rolling a six-sided die result in the state space of $\Omega = \{1, 2, 3, 4, 5, 6\}$, but it is less plausible to think that an agent knows every possibility of more complicated events, for instance, say, the possible outcomes of a bank merger or the possible outcomes of a sophisticated science experiment (that is, sometimes we are unaware of some possibilities). A third way to interpret (K0″) is that agents know all tautologies or logical truisms. As I mentioned above, there are often several ways to interpret the same event. Consider the event Ω. Since a logical truism is true in every state (by definition), and Ω obtains when any state ω obtains, we can identify Ω as equivalent to any logical truism. Surely, there is a sense that realistic agents do not know every logical truism. I address this issue in Chapter 3 and Chapter 5. Nonetheless, it is straightforward to show that (K0″) also follows from the set-theoretic definition of knowledge without any assumptions on P. By the set-theoretic definition of knowledge $\omega \in K(\Omega)$ if and only if $P(\omega) \subseteq \Omega$, but this follows from the definition of P (that is, P is a non-empty subset of states

from Ω), and is true at every state ω'. So $\omega \in \Omega$.

1.3.1 Restrictions on P, and their Consequences

So far, I've said little about how an agent's possibility function is to be structured. It should not be surprising that modifications to P have implications for the knowledge operator K (after all, knowledge is defined in terms of P). Consider three restrictions on P:

(P1): $\omega \in P(\omega)$.

(P2): If $\omega' \in P(\omega)$, then $P(\omega') \subseteq P(\omega)$.

(P3): If $\omega' \in P(\omega)$, then $P(\omega) \subseteq P(\omega')$.

Each of the restrictions (P1)–(P3) have distinct implications for K. Of course, there are many other restrictions on P to consider. But (P1)–(P3) have important epistemic consequences, so I will focus attention on them.

First, consider (P1). This restriction can be understood as requiring that the agent always considers the actual state as possible (recall that $P(\omega)$ returns the set of epistemically possible states for S when the actual state is ω). Lewis (1996) endorses a similar constraint on epistemic possibility, one he calls "the rule of actuality".[7] It is easy to show that this restriction provides an essential property on knowledge: factivity. The factivity of knowledge can be represented with the following:

(K1): $K(E) \subseteq E$.

If (K1) is true, then if an agent knows E, then E is true. Next, I will show that (P1) entails (K1). Assume that $\omega \in P(\omega)$ and that $\omega \in K(E)$. By the set-theoretic definition of knowledge, $P(\omega) \subseteq E$. But, by (P1), $\omega \in P(\omega)$, so $\omega \in E$. I take

[7] See Lewis (1996), p. 554. As Lewis remarks, "the possibility that actually obtains is never properly ignored; actuality is always a relevant alternative...".

the factivity of knowledge, and correspondingly (K1), to be a necessary condition on knowledge. Any model of knowledge without this feature is not a model of knowledge.

The remaining two restrictions on P provide for various iterations of knowledge. As I explained above, "$K(E)$" is an event (it is the event "the agent knows E"), so it is possible for an agent to know or not to know this event. Consider the following:

(K2): $K(E) \subseteq K(K(E))$.

This principle is the familiar KK principle, and can be read "if an agent knows E then she knows that she knows E". It is controversial whether (K2) is true for all or any epistemic agents. I discuss arguments for and against this principle in Chapter 4. As it turns out, the restriction (P2) on P entails (K2). To see why, assume (P2) and assume that $\omega \in K(E)$. To show that $\omega \in K(K(E))$ it is enough to show that $P(\omega) \subseteq K(E)$. Assume that some $\omega' \in P(\omega)$. By (P2), this means that $P(\omega') \subseteq P(\omega)$. But because $\omega \in K(E)$ this means that $P(\omega) \subseteq E$, and hence, $P(\omega') \subseteq E$. So it must be that $\omega' \in K(E)$.

(K2) is sometimes referred to as the "positive introspection" axiom for knowledge: agents know that they know. The following "negative introspection" axiom characterizes agents who know that they don't know:

(K3): $\neg K(E) \subseteq K(\neg K(E))$.[8]

Again, read (K3) as "if an agent doesn't know E then she knows that she doesn't know E". As should be obvious, (K3) is highly controversial for realistic, non-ideal agents. I discuss this principle in Chapter 4.

[8]Where "$\neg K(E)$" is given by the following: $\neg K(E) = \{\omega : P(\omega) \not\subseteq E\}$, and "$\neg$" is hence interpreted as the set-theoretic operation of complementation "$-$". Intuitively, an agent doesn't know some event if her possibility function is such that it is not "contained" in E—in this case there are non-E states that are epistemically possible for the agent, so she cannot rule out the event "$\neg E$".

Finally, it is important to note that (K1) and (K3) entail (K2). By (K1) and (K3), $\neg K(E) = K(\neg K(E))$, and hence $K(E) = \neg K(\neg K(E))$. Substituting $\neg K(E)$ for E, it follows that $K(\neg K(E)) = \neg K(\neg K(\neg K(E)))$. By a chain of equalities we get, $K(E) = \neg K(\neg K(E)) = K(\neg K(\neg K(E))) = K(K(E))$.

1.3.2 Events and Propositions

In the set-theoretic model of knowledge, agents know events. But epistemologists standardly discuss knowledge of propositions. There is a straightforward relationship between events and propositions. This relationship goes some way toward showing how the set-theoretic model of knowledge is formally equivalent to more familiar logics of knowledge (epistemic logics).

The set-theoretic model of knowledge I've given above does not contain an object language. This is one reason why there is some indeterminacy about the interpretation of events. Yet, in this way, the set-theoretic model of knowledge bears some close resemblance to models of probability. Standard models of probability determine probabilities for events, understood as collections of "basic outcomes" (which are states). For example, consider the random experiment of rolling a six-sided die. The possible states are, again, $\Omega = \{1, 2, 3, 4, 5, 6\}$. If we assume that the die is a fair die, then each state is assigned probability $1/6$. An event E is a collection of states (or basic outcomes), and its probability is determined by the probability calculus. These models do not determine probabilities for "language like" structures such as formulas in propositional logic, for instance.

Nevertheless, there is a way to translate between "language like" formulas and states. The basic idea is that a state ω, as a "full description of the world", is a set of all formulas true at that state. I next survey a possible-worlds model of knowledge (making precise the notion of a formula) and show how this model is formally equivalent to the set-theoretic model of knowledge.

1.4 A Possible-Worlds Model of Knowledge

The essential difference between the possible-worlds model of knowledge and the set-theoretic model of knowledge is that the former has an object language. Begin with a set of atomic propositions Φ, and close off the set under the familiar logical relations of $\neg$, $\wedge$, $\vee$, and $\rightarrow$, and the modal operator "K".[9] The closure of Φ under these logical relations is the object language.

Call a Kripke structure M (over Φ) a tuple $(S, \pi, \mathcal{K})$. Formally, S is a set of states identical in logical kind to the set of states Ω in set-theoretic information structures from the previous section. I refer to states in Kripke structures with 'S' and states in information structures with 'Ω' to help distinguish the two models. Next, π is an "interpretation function" that assigns a truth value to members of Φ at each state. That is, $\pi(s) : \Phi \rightarrow \{$ **true, false** $\}$ for each $s \in S$. $\mathcal{K}$ is a binary relation on S, similar in interpretation to the possibility function P in information structures.

The semantics for the familiar logical relations are standard.

(i) $(M, s) \models p$ (for atomic proposition $p \in \Phi$) if and only if $\pi(s)(p) =$ **true**.

(ii) $(M, s) \models \varphi \wedge \psi$ if and only if $(M, s) \models \varphi$ and $(M, s) \models \psi$.

(iii) $(M, s) \models \neg\varphi$ if and only if $(M, s) \not\models \varphi$.

Clauses for '$\varphi \vee \psi$' and '$\varphi \rightarrow \psi$' are unnecessary, because '$\varphi \vee \psi$' is defined as '$\neg(\neg\varphi \wedge \neg\psi)$' and '$\varphi \rightarrow \psi$' is defined as '$\neg(\varphi \wedge \neg\psi)$'.

The "epistemic" clause concerns the operator K. In information structures, an event E is known if the agent excludes all non-E states. This is the same idea for the possible-worlds model, though I am now able to frame this idea in terms of truth, because we have an object language.

[9] My presentation and discussion of the possible-worlds model of knowledge is adapted from Fagin et al. (1995) and Rubinstein (1998).

(iv) $(M, s) \models K\varphi$ if and only if $(M, t) \models \varphi$ for all t such that $(s, t) \in \mathcal{K}$.

Clause (iv) can be read as "φ is known at state s if and only if φ is true in all states accessible from s". Or, as this is sometimes condensed, "knowledge is truth in all possible states". Because many philosophers prefer the nomenclature of "worlds" rather than "states", the idea is that "knowledge is truth in all possible worlds".

1.4.1 The Relationship Between Information Structures and Kripke Structures

As should be clear, information structures and Kripke structures bear close similarities. Technically, information structures are what are called "Kripke frames", the tuple $(S, \mathcal{K})$ *without* the interpretation function π. This is because information structures do not have an object language. Once this difference is accounted for, information structures and Kripke structures are formally equivalent.

Because there are no propositions in information structures, define the event "p is true" as the set of states s where p is true. That is, $e_p^M = \{s : (M, s) \models p\}$. It is then straightforward to define an event $\mathsf{ev}_M(\varphi)$ for each $\varphi \in \Phi$.

(i) $\mathsf{ev}_M(p) = e_p^M$ (when p is atomic).

(ii) $\mathsf{ev}_M(\varphi \wedge \psi) = \mathsf{ev}_M(\varphi) \cap \mathsf{ev}_M(\psi)$.

(iii) $\mathsf{ev}_M(\neg\varphi) = \Omega - \mathsf{ev}_M(\varphi)$.

(iv) $\mathsf{ev}_M(K\varphi) = K(\mathsf{ev}_M(\varphi))$.

The intuitive identification of events with formulas in Φ results from an understanding that $\mathsf{ev}_M(\varphi)$ is intended to represent the event "that φ holds".

Given the above translation scheme, one may translate between Kripke structures and information structures without loss of logical content. Fagin et al. (1995) sketch a proof of the formal equivalence of these two models.[10]

[10] See Fagin et al. (1995), pp. 39–40.

1.4.2 Advantages

Even though Kripke structures and information structures can be viewed as equivalent models of knowledge, each have interpretational advantages over the other for discussion of various issues.

In Chapter 4 I discuss arguments for the KK principle (principle (K2) from page 22), and, as it turns out, the set-theoretic model of knowledge is more helpful than the possible-worlds model. In that chapter I address why the set-theoretic model is more helpful, but the simple reason is that information structures allow for an intuitive discussion about the structure of the possibility relation when we aim to model the knowledge of cognitively bounded agents such as humans. In information structures the possibility function is a relation between states, *without* specific reference to propositions or language. As I'll show in Chapter 4, this makes the task of modeling various cognitive phenomena easier.

In Chapter 3 I discuss issues of epistemic closure (especially property (K0) from page 19 above). There, I make explicit reference to the syntactical component of our knowledge and, hence, favor the possible-worlds model of knowledge. Nothing of logical substance is lost between framing an issue in terms of one model over the other. In Chapter 5 I give a general discussion about interpretive issues of epistemic logic.

CHAPTER 2

COMMON KNOWLEDGE, COORDINATION, AND CONTEXT

2.1 Introduction

Many have argued, and it is intuitively plausible, that various social coordination problems require epistemic sophistication in the form of common knowledge. But, under scrutiny, common knowledge also seems unattainable. Such observations work to form a paradox, largely unnoticed in the philosophical literature. I will give an analysis of the paradox and offer a solution. In particular, I will argue that rational coordination *does not* require common knowledge, contrary to our intuition. Next, I will show how my solution informs a recent debate in epistemology concerning the relationship between knowledge and action. I will offer a challenge to John Hawthorne and Jason Stanley's view that it is rationally appropriate to treat a proposition as a reason for acting only if the proposition is known—I will show that there are ordinary cases of rational coordination where agents lack knowledge, though are still rational to act.

2.2 The Importance of Higher-Order Knowledge

Common knowledge, or something like it, appears indispensable for understanding much of our social behavior. Call a proposition p common knowledge (for a group $\mathcal{G}$) when everyone knows p, everyone knows that everyone knows p, everyone knows that everyone knows that everyone knows p, ... *ad infinitum*.[1] Further, say that

[1]Note that 'common knowledge', as I've defined it, is a technical expression. My use is to be distinguished from the ordinary use of 'common knowledge', which typically means "something known by most people".

everyone knows p if and only if every person g in the relevant group $\mathcal{G}$ knows p.

Common knowledge may seem like a rarefied technical concept, but it has proven useful in many philosophical domains: David Lewis (1969) famously argued that conventions (either in language or action) must be common knowledge, Jane Heal (1978) has argued that common knowledge can justify action in a way that weaker notions cannot, Robert Stalnaker (1975) employs common knowledge to explicate the common ground (shared background information) in a conversation, and many game theoretic models of interactive decision making (Fundenberg and Tirole (1991), for example) assume that players have common knowledge of the rules of the game and the utility functions of the other players.

In *Rational Ritual*, Michael Chwe (2001) has argued that many of our public social practices are best understood as principally generating common knowledge. For example, consider the prevalence of inward-facing circles at meeting places. Much has been made of King Arthur's Round Table for its symbolism of equal social status, but round tables also seem well suited to generate common knowledge. For, at a round table, it is easy for each member to see that every other member is paying attention, and also that she (the observer) is likewise being observed. These meeting places are not a modern invention—throughout much of the desert southwest United States, ancient circular structures called kivas are found. Chwe writes that "most interpreters see the function of kivas... as ritual structures for the villages, where public activities could be held".[2] Many kivas are below ground and took considerable effort and resources to construct. Why would tribes like that of the Pueblo people expend great effort in building these complicated structures? One explanation is that these public structures promote community and collective identity. Surely this is correct. But Chwe thinks that these structures also have a distinctly epistemic benefit. Perhaps, by situating members where everyone can observe everyone else simultaneously, important community announcements can be

[2]Chwe (2001), p. 30.

made public and become common knowledge.

Many repressive regimes appear to understand the value of public information and the common knowledge it seems to bring. Pubic meeting places, public channels for information, and public symbols are often the first targets by dictators when there is unrest. The rebels, in contrast, often compete eagerly for such public forums.

Why might common knowledge be important for social change? A rebellion against a regime is what is called a "coordination problem": each member of a rebellion wants to show up at a demonstration or attempt an overthrow if and only if the others do as well. Showing up at a demonstration alone is costly, often fatal. In this sense, a coordination problem is one where everyone's interests coincide. So, I want to show up at the demonstration against the dictator only if you do, and vice versa. Common knowledge then seems essential for our coordination because I will show up at the demonstration only if I *know* you will too. That is, if I don't know that you will show up then I won't show up. And likewise for you. But, it seems, I also need to know that you know that I will show up—for, if I don't know whether you know that I will show up, it seems likely that I won't show up. If a symbol, or graffiti, or a poster, or a message is public, then we all know that we all know that... a demonstration will take place. Absent public information, it seems less likely that there will be common knowledge of a demonstration, and, hence, it seems less rational to participate in a demonstration because there is uncertainty about what others will do.

To further illustrate the connection between common knowledge and coordination, consider the following problem:

> *Coordinated Attack*: Two divisions of an army, each commanded by a general, are camped on two hilltops overlooking a valley. In the valley awaits the enemy. It is clear that if both divisions attack the enemy simultaneously they will win the battle, while if only one division attacks it will be defeated. As a result, neither general will attack unless he is

> absolutely sure that the other will attack with him. The commanding general of the first division wishes to coordinate a simultaneous attack (at some time next day). The generals can communicate only by means of messengers. Normally, it takes a messenger one hour to get from one encampment to the other. However, it is possible that he will get lost in the dark or, worse yet, be captured by the enemy. Fortunately, on this particular night, everything goes smoothly. How long will it take for them to coordinate an attack?[3]

The intuitive answer to Coordinated Attack is that the two generals will never be able to rationally coordinate an attack. Call the two generals A and B. Suppose A sends B the message "attack at dawn". Should B attack? It seems not. For, while the message was sent successfully from A to B, A doesn't know that B received the message (after all, the messenger could have been captured). Correspondingly, B knows that A doesn't know that B received the message, and so B won't attack. But, B could send a reply to A with the acknowledgement "message received". Would A attack in this case? Again, it seems no. For, now, B doesn't know that A received the acknowledgement. Even if B sent the message "I will send no more messages", B wouldn't know if A received the message.

The intuition is, then, that no finite sequence of successful deliveries of acknowledgement messages is enough for A and B to rationally coordinate their attack. From the example, it is clear what would work: if it were common knowledge between A and B that they attack at dawn, then A and B will attack at dawn. This is because common knowledge serves to erase the uncertainty between the two generals about whether the other has received the appropriate message. In this case, A knows that B knows, B knows that A knows, A knows that B knows that A knows, ... etc.

[3]Wording from Fagin et al. (1995), p. 176–177. Fagin et al. remark that a version of the coordinated attack problem was first introduced by the IBM researcher Gray (1978).

2.3 The Paradox

In the preceding section I've described how things *seem*. It seems that coordination problems like Coordinated Attack require common knowledge. And, as I'll argue below, it seems that we often rationally solve coordination problems (similar in structure to Coordinated Attack). But, focusing on the complexity of common knowledge (everyone knows that everyone knows that...), it also seems that common knowledge is unattainable for real, non-ideal agents. Of course, we often know that someone knows something. But for a proposition to be common knowledge there must be an *infinite* iteration of my knowing that you know that I know something. Could we ever have such knowledge?

Together these three premises are the following:

(1) We often rationally solve coordination problems.

(2) Rationally solving coordination problems requires common knowledge.

(3) Common knowledge is unattainable.

Each of the premises (1)–(3) are plausible (or so I'll argue), yet the collection is jointly inconsistent: one premise must be false.[4] In this section I'll motivate each premise and explain why it is plausibly true. In the following section I will resolve the paradox. I will then show how my solution informs a recent debate in epistemology concerning the relationship between knowledge and action.

2.3.1 We Often Rationally Solve Coordination Problems

A coordination problem is an interactive decision where the agents' interests coincide (with respect to the decision at hand). Thomas Schelling's (1960) example involves

[4]The common knowledge paradox was first formulated by the computer scientists Halpern and Moses (1990).

a man and his wife who are separated and attempt to find each other in a large department store. Each one doesn't care where they look (aisle 1 or aisle 2?) only that they find each other. Call the couple Amy and Bobby. If Amy and Bobby both choose to look in the same aisle, they receive a positive payoff. If Amy and Bobby's choices do not match, they receive no payoff (they remain separated). This game is a coordination game because it can be seen from the payoffs that Amy and Bobby don't really care which aisle they choose, as long as they pick the same aisle. From their perspective their interests coincide. That is, the only way to "win" is for both players to win.

Coordination problems are common to social life. When I attempt to meet a philosopher to talk about a paper, I'm trying to solve a coordination problem. It doesn't matter where we meet (at her office or at a coffee shop or at the library), so long as we both show up at the same place.[5] David Lewis (1969) recognized that we often solve coordination problems by means of convention. To take an obvious example, society faces a coordination problem when trying to organize traffic on the street. It is better for everyone that we (collectively) pick a side to drive on, and it doesn't really matter which we pick. It is a convention in the United States to drive on the right and it is a convention in England to drive on the left. We have conventions about money (we could have used silver instead of gold while we were on the gold standard), we have conventions about measurement (the United States could have used something similar to the metric system), and we have conventions about language (the word 'car' denotes things we drive around in, though we could have referred to them with the word 'bolley').

There are also many mundane examples of coordination problems that we regularly face in daily life. When a couple decides where to meet for dinner after work,

[5]Of course, matters aren't *always* so simple. The philosopher might pick a location that allows her to accomplish other goals (like checking email at her office). In that respect, it might be rational for her to go to the office regardless of whether I show up.

they face a coordination problem. Scheduling is often a coordination problem—for most purposes, it doesn't matter whether the colloquium is in room 311 or 312, provided everyone shows up in the same place. In general, coordination problems arise when groups of agents have common goals that require everyone taking the same action. We face such problems routinely.

Next, I think we rationally solve many of these coordination problems. We make dinner plans, we attend meetings, we meet to discuss philosophy papers. Such behavior is usually purposeful and successful. We don't accidentally stumble on our meeting locations, we arrive as intended and with reason for our action. These remarks, I suggest, are enough to make the first premise plausible: we often rationally solve coordination problems.

2.3.2 Rationally Solving Coordination Problems Requires Common Knowledge

I gave premise (2) above in informal terms: the idea is that rational coordination necessitates the existence of common knowledge. A more precise formulation is the following:

> (2.1) The rational solution R (based on p) to a coordination problem has it, as a necessary condition, that p is common knowledge.

There are several technical results in the philosophical logic literature and the economic theory literature that support premise (2.1). Fagin et al. (1995) show that no finite amount of communication between the two generals is enough to ensure rational coordination.[6] Ariel Rubinstein's (1989) electronic mail game shows, similarly, that approximate common knowledge cannot establish coordination.[7] However, I

[6] See especially Fagin et al. (1995), p. 182: "an attack is never attainable in *any* run of *any* deterministic protocol for coordinated attack. Thus, even if every message is delivered, coordinated attack is not possible, as long as there is the possibility that messages will not be delivered" (original emphasis).

[7] See also Osborne and Rubinstein (1994), pp. 81–84.

am less interested in the technical results and more interested in providing intuitive grounds to support the premise.

From the Coordinated Attack problem, surely it is possible that the two generals happen to arrive at the valley of the enemy at dawn by coincidence (supposing no messages were sent). But such an outcome wouldn't be rational. A ranking officer could criticize the generals in the following way: "you (general A) didn't know that general B knew to attack at dawn—you shouldn't have attacked at dawn".

The ranking officer's remarks sound correct to me, and they support premise (2.1). To see this clearly (and in anticipation of the discussion of premise (3) below), consider the contrapositive of premise (2.1):

> (2.2) If one does not know that other player knows some part of the common knowledge sequence for p, then it is not rational to coordinate with solution R based on p.

To understand (2.2), I need to define the "common knowledge sequence".[8] For two agents A and B, the common knowledge sequence when p is common knowledge is the following:

> (ck1) A knows p.
>
> (ck2) B knows p.
>
> (ck3) A knows A knows p.
>
> (ck4) A knows B knows p.
>
> (ck5) B knows B knows p.
>
> (ck6) B knows A knows p.
>
> $\vdots$
>
> *ad infinitum*.

[8]The common knowledge sequence is also referred to as the "common knowledge hierarchy".

The conjunction of this common knowledge sequence is equivalent to "everyone (A and B) knows that p, everyone knows that everyone knows p, everyone knows that everyone knows that everyone knows p, ... *ad infinitum*"—it is merely an unpacking of the definition of common knowledge.

When the ranking general admonishes general A above, she cites a member from the common knowledge sequence: because general A didn't know that general B knew to attack at dawn, general A shouldn't have attacked at dawn. And such admonition may come from any member of the common knowledge sequence, if an agent doesn't know that member of the sequence. The intuition is this: if an agent doesn't know a member of the common knowledge sequence then she cannot predict the behavior of the other player and cannot rationally play the coordinating strategy. If I don't know whether you know that I know to attend the demonstration, I cannot predict that you will attend, and so I won't attend (for fear of arrest, say). Likewise, if you cannot resolve whether I know this about you, you cannot predict what I will do, and so you won't attend.

Premise (2.2) is equivalent to requiring common knowledge for rational coordination. Again, the contrapositive of (2.1) is that if p is not common knowledge then it is not rational to take R as a solution to the coordination problem. So, p is not common knowledge if any member of the common knowledge sequence is not known. Hence, I suggest, premise (2.2) and thereby premise (2) is plausible.

2.3.3 Common Knowledge is Unattainable

Yet, premise (2.2) is supposed to sound demanding: common knowledge seems informationally complex. To rationally coordinate with another agent based on the proposition p (say, "we attack at dawn"), I need to know a lot. I need to know the proposition, I need to know that the other player knows the proposition, I need to know that the other player knows that I know the proposition, I need to know that the other player knows that I know that the other player knows that I know the

proposition, and so on. And these propositions are actually toward the beginning of the common knowledge sequence. That is, these propositions are relatively easy, given that the sequence extends infinitely. It seems doubtful that I could ever know *all* these propositions.[9]

Another way to motivate the third premise is to note our human limits on parsing sentences. At some point, members of the common knowledge sequence become meaningless. For example, it is exceedingly difficult to interpret the meaning of the following:

> (4) $S1$ knows that $S1$ knows that $S2$ knows that $S2$ knows that $S1$ knows that $S2$ knows that $S1$ knows that $S2$ knows that $S1$ knows that $S2$ knows that $S1$ knows that $S2$ knows that $S1$ knows that $S2$ knows that $S2$ knows that $S2$ knows that $S1$ knows that $S1$ knows that p.

I don't understand proposition (4)—at some point in reading the sentence I lose track of who knows what. And because I don't understand proposition (4), it seems beyond my ken to know whether it is true. Most of the members of the common knowledge sequence will be more complex than proposition (4). Hence, I suggest, it is plausible that common knowledge is unattainable.

Some philosophers find it intuitive that in ordinary cases of public announcement, agents have common knowledge.[10] Suppose Abe and Beth are sitting next to each other in a quiet room when Beth says to Abe, "it is raining" (and suppose it is raining).[11] Should we say that it is common knowledge for Abe and Beth that it is

[9]Paternotte (2011) suggests, "...agents have neither the memory capacity to stock an infinite number of statements nor the time to infer every one of them". I'm not convinced that citing storage issues is the best way to defend premise (3) (see my discussion of finitary models below), but it is a concern raised in the literature.

[10]See Vanderschraaf and Sillari (2009) and also Paternotte (2011).

[11]Suppose, further, that Abe has no reason to think that Beth is not reliable and that Abe and Beth make eye contact and that they have no reason to think that the other is distracted, etc.

raining? First, note that 'common knowledge' is a technical expression. We cannot have intuitions about the expression *itself*, so we must think about the definition. Should we say that Abe and Beth know that it is raining? Yes. Should we say that Abe knows that Beth knows that it is raining? Again, yes. But should we say that Abe knows that Beth knows that Abe knows that Beth knows that Abe knows that Abe knows that Beth knows that Beth knows that Abe knows that Abe knows that Abe knows that Beth knows that Beth knows that Abe knows that Beth knows that Abe knows that Abe knows that Beth knows that Abe knows that Beth knows that Abe knows that Abe knows that Beth knows that Abe knows that Beth knows that Abe knows that it is raining? No. I don't have an intuition whether Abe knows *this* because I don't understand the proposition in question. So, it can't be the case that it is intuitive that Abe and Beth have common knowledge, even in this simple case. One may have other reasons to think common knowledge is possible, but it cannot be supported by intuition alone.

Is Common Knowledge Possible?

The case for common knowledge cannot rest only with intuitions. What other reasons might be given in defense of common knowledge? As it turns out, there are finitary models of common knowledge that are logically equivalent to the common knowledge sequence. Might this help the case for common knowledge?

In general, a finitary model of common knowledge collapses the infinite common knowledge sequence into some finite proposition, some finite piece of information, or a finite procedure for checking whether common knowledge obtains. If we assume that our language is finite and that there are only a finite number of possible states of the world (both reasonable assumptions for non-ideal agents), finitary models of common knowledge have the following property for K, the knowledge operator: there is a finite n such that $K^n_{\mathcal{G}}p = K^\infty_{\mathcal{G}}p$, where $K^n_{\mathcal{G}}p$ represents a finite collection of members (up to the n^{th} member) from the common knowledge sequence and $K^\infty_{\mathcal{G}}p$

represents the infinite common knowledge sequence.[12] Such a result might make it seem plausible that real, non-ideal human agents could possess common knowledge, for all that is really needed for common knowledge is knowledge of a finite number of members from the common knowledge sequence. However, I will argue that finitary models do not show that common knowledge is possible for real, non-ideal agents. Note at the outset that while n may be finite, it may still be unboundedly large.

There are several finitary models of common knowledge that have been proposed in the literature.[13] Robert Aumann (1976) is credited with the first such model, but he makes several idealizing assumptions that are inappropriate for real, non-ideal agents.[14] In its place I will consider the model given by Rubinstein (1998): this model can be shown to be a generalization of Aumann's model.

I introduced a model of knowledge (a "set-theoretic" model of knowledge) in Chapter 1. The model of common knowledge I discuss below is built from this model. I next review several important features of the set-theoretic model of knowledge.[15]

An information structure (Ω, P) is a pair with a set of states Ω, and a possibility function P. A state ω is a "full description of the world", which resolves all matters of fact for the problem at hand. States are mutually exclusive. For example, when considering the outcome of rolling a six-sided die, the states are $S = \{1, 2, 3, 4, 5, 6\}$. Possibility functions, often indexed P_i for each agent i, return for each $\omega \in \Omega$, all states ω' that are epistemically possible for i at ω. So, if $P_i(\omega) = \{\omega, \omega'\}$, then both ω and ω' are epistemically possible for i at ω.

[12] See Fundenberg and Tirole (1991), p. 544.

[13] For brevity, I omit discussion of fixed-point analyses of common knowledge.

[14] Aumann assumes that agents are able to partition their information space, which entails that agents reason according to the positive and negative introspection axioms. These axioms are controversial: I argue against the KK principle (positive introspection) in Chapter 4. As well, I further discuss partitional information spaces and their epistemic implications in Chapter 5.

[15] Set-theoretical models of knowledge are equivalent to Kripke frames. See Fagin et al. (1995) for further discussion. Such models do not make explicit reference to an object language, and in this regard are akin to models in probability theory.

Just as one may make restrictions on the "accessibility relation" in epistemic logics, one may make restrictions on possibility functions. The only restriction I make on P is the following:

(P1): $\omega \in P(\omega)$.

The interpretation of (P1) is that agents never exclude the actual state as epistemically possible. It will turn out that (P1) ensures that the knowledge operator or function K, introduced next, is factive. I take the factivity of knowledge as a minimal requirement for any model of knowledge.

Call any collection E of states ω "the event E". Say that the event E is known by i at ω (knowledge at a state) if and only if $P_i(\omega) \subseteq E$. The idea behind this definition is similar in spirit to Hintikka (1962) and relevant alternatives theories of knowledge: to know p is to exclude all non-p worlds or states as epistemically possible. Next, define a knowledge operator, induced from knowledge at a state, with the set $K_i(E) = \{\omega : P_i(\omega) \subseteq E)\}$. $K_i(E)$ is interpreted to mean "the decision maker i knows E". If (P1) holds for the operator K, then K has the following property (the factivity of knowledge):

(K1): $K(E) \subseteq E$.

Two more definitions. An event E is *common knowledge* between agents 1 and 2 at the state ω if and only if ω is a member of all sets of type $K_1(E), K_2(E), K_1(K_2(E)), K_2(K_1(E))$, and so forth (the preceding is read "agent 1 knows E, agent 2 knows E, agent 1 knows that agent 2 knows E, agent 2 knows that agent 1 knows E).[16] Next, call an event E *self-evident* for P_1 and P_2 if for all $\omega \in E$ and for both i, $P_i(\omega) \subseteq E$. The interpretation is that a self-evident event is such that whenever it occurs, agent 1 and agent 2 knows that it occurs.

[16]See Chapter 1 for a discussion on translating between events and propositions.

There are three important properties of the above common knowledge model that I wish to highlight. First, in the model, a proposition p is common knowledge if and only if everyone knows p, everyone knows that everyone knows p, ... *ad infinitum*. So, common knowledge has the familiar rendering we would expect. Second, it can be shown that this model is equivalent to standard models of common knowledge framed in terms of epistemic logic.[17] Third, Rubinstein's model has the following finitary property:

> **Theorem 2.1** (Finitary Common Knowledge): Assume that P_1 and P_2 are information structures satisfying (P1). Let K_1 and K_2 be the knowledge operators induced from P_1 and P_2, respectively. The event $E^\star$ is common knowledge at ω if and only if it includes a self-evident event E containing ω.
>
> **Proof**: Assume that there is a self-evident event E so that $\omega \in E$ and $E \subseteq E^\star$. By the definition of a self-evident event, for both i, $E \subseteq K_i(E)$ and, by P1, $K_i(E) \subseteq E$; thus $K_i(E) = E$ for both i, which implies that $K_iK_j \ldots K_i(E) = K_iK_j \ldots K_j(E) = E$. By (K0), since $E \subseteq E^\star$ we have $K_i(E) \subseteq K_i(E^\star)$ and thus $E = K_iK_j \ldots K_i(E) \subseteq K_iK_j \ldots K_i(E^\star)$ and $E = K_iK_j \ldots K_j(E) \subseteq K_iK_j \ldots K_j(E^\star)$. So, since $\omega \in E$, ω is a member of all sets of the type $K_iK_j \ldots K_i(E^\star)$ and $K_iK_j \ldots K_j(E^\star)$. That is, $E^\star$ is common knowledge at ω.
>
> For the other direction, if $E^\star$ is common knowledge at ω, take E to be the intersection of all sets of the type $K_iK_j \ldots K_i(E^\star)$ and $K_iK_j \ldots K_j(E^\star)$. Because $E^\star$ is common knowledge at ω, $\omega \in E$. By (K1), which follows from (P1), $E \subseteq E^\star$. To show that E is a self-evident event, one just has to verify that for any $\omega \in E$, $P_i(\omega) \subseteq E$. This follows from the fact that because $\omega \in E$, ω belongs to any set of the form $K_iK_j \ldots K_i(E^\star)$.[18] ■

[17] See Fagin et al. (1995).

[18] Proof adapted from Rubinstein (1998), p. 58–59. (K0) is a closure property on K. (K0): if

Finitary Common Knowledge shows that whether an event is common knowledge, whether the infinite common knowledge sequence for a group $\mathcal{G}$ is true, is equivalent to an existence claim about a self-evident event. Hence, the question about the possibility of common knowledge is effectively a question about the existence of self-evident events with the appropriate entailment property. Are there any such self-evident events?

Below I will argue in the negative: for non-trivial propositions of interest, there never exist self-evident events with the appropriate entailment property. First, note the epistemic demands of self-evident events. From the definition, an event is self-evident if it is known whenever it is true. Self-evident events seem similar to an important category of proposition that have occupied epistemologists, but they are actually even more demanding than these traditionally discussed propositions. Epistemic foundationalists have searched for a class of propositions, we may call them "luminous propositions", that we are in a position to know whenever they obtain.[19] In this sense, luminous propositions are never "hidden" from us. As Williamson (2000) describes him, Descartes' project in the *Meditations* was to find such a luminous realm in the mental. Similarly, McDowell (1989) conceived of pain as luminous: if one is in pain then one is always in a position to know that one is in pain. Though the existence of luminous propositions is still very much contested in epistemology (Williamson (2000) gives a forceful argument that there are none), I wish to point out that self-evident events are actually logically stronger than luminous events.

$E \subseteq F$ then $K(E) \subseteq K(F)$. See Chapter 1 for a further discussion of (K0). I discuss general closure principles in Chapter 3.

[19]See Williamson (2000), especially Chapter 4. As he defines them, a condition C is luminous if and only if for every case α, if in α C obtains, then in α one is in a position to know that C obtains. Further, Williamson says "to be in a position to know p, it is neither necessary to know p nor sufficient to be physically and psychologically capable of knowing p... If one is in a position to know p, and one has done what one is in a position to do to decide whether p is true, then one does know p", p. 95.

That is, if $p^\star$ is self-evident then it is luminous, but the converse is not true. One might think that it is possible to be in pain while not being in a position to know that one is in pain. Or one might think that it is possible to be appeared to as if q but not be in a position to know that one is actually appeared to as if q. Because of the above entailment relation, reasons to deny the existence of luminous events are reasons to deny the existence of self-evident events. So, whatever doubts one may have about luminous events, doubts should be magnified for self-evident events, and, correspondingly, common knowledge.

My second argument against the existence of self-evident events concerns the above case of Abe and Beth. Again, consider Abe and Beth sitting next to each other in a quiet room. Beth says to Abe, "it is raining" (and suppose it is raining).[20] Clearly, if there ever is a case of common knowledge this would be it. By Finitary Common Knowledge, if the event "it is raining" is common knowledge for Abe and Beth, then there exists some self-evident event.

What are the candidate events that might be self-evident for Abe and Beth, and thereby support common knowledge? Here are two initial possibilities: the event that it is raining E_R and the event that Beth says that it is raining E_S. First, the event that Beth says that it is raining *cannot* serve as the self-evident event for Abe and Beth, because E_S does not entail E_R (which is required by Finitary Common Knowledge). This is because there are ways for E_S to obtain without E_R obtaining—for instance, it is possible that Beth is mistaken when she tells Abe that it is raining. Instead, might we have reason to think that the event that it is raining, E_R, is self-evident for Abe and Beth? I suggest, no. The reason we should think that many standard empirical or contingent events like "it is raining" are not self-evident is because we think there can be "evidential breaks" between the event happening and our knowing that it happened. The event "it is raining"

[20] Again, suppose that Abe has no reason to think that Beth is not reliable and that Abe and Beth make eye contact and that they have no reason to think that the other is distracted, etc.

is not self-evident for Abe because one way for the event to obtain is when it is raining and Abe is camped in the windowless basement of the library reviewing old philosophy journals. In such a case there would be a "break" between his evidence or information and the event (i.e., there is a way for the event to obtain without Abe receiving any information that it obtains). Note that the definition of a self-evident event is *modal* in character. Again, an event E is self-evident for S if and only if *any way* E obtains is such that S knows E. But we think there are ways for it to rain without our knowing that it rains (the above basement case provides one such example). So the event that it is raining E_R is not self-evident for Abe, hence it cannot underwrite a claim of common knowledge.

What other candidate event might be self-evident for Abe and Beth? As it turns out, a careful study of the proof of Finitary Common Knowledge provides a suggestion. From the proof, we need some self-evident event E such that $E \subseteq E^R$. But because knowledge is factive, if $E = K(E^R)$, then E would have the correct entailment property (because $K(E^R) \subseteq E^R$).[21] Might the event "Abe knows that it is raining" serve as a self-evident event that supports common knowledge that it is raining? There are two reasons to answer in the negative. First, similar to the above, knowing an event appears, at least on the surface, to closely resemble an empirical or contingent event. I previously argued that with canonical empirical/contingent events such as "it is raining" there are ways for the event to obtain *without* the agent knowing that it obtains. Because an event of knowing seems to resemble an empirical or contingent event, we should think that there are ways to know *without* the agent knowing that she knows.[22] Second, if the event "Abe knows that it is raining" is self-evident for Abe *and* Beth, this means that in any way Abe comes to know that it is raining, Beth knows that Abe knows this. But, surely, on any realistic model of knowledge, there will be states such that Abe knows it is raining

[21]Recall that the factivity of knowledge is equivalent to the following, for any E: $K(E) \subseteq E$.

[22]I discuss this claim and the related KK principle in Chapter 4.

but Beth doesn't know that Abe knows this. As one such example, suppose Abe is outside while Beth is camped in the library basement reviewing old philosophy journals. When it rains, Abe knows that it is raining, but Beth fails to notice the rain (there are no windows in the basement) and does not know that Abe knows that it is raining.[23] It is easy to generate similar examples. So we have reason to think that the event "Abe knows it is raining" is not self-evident for Abe and Beth.

This second point is worth considering carefully. From Finitary Common Knowledge, if an event such as "it is raining" is common knowledge for Abe and Beth, then there must exist some self-evident event E such that E entails the event "it is raining". Letting E be the event "Abe knows that it is raining", it is clear that E satisfies the required entailment relation, because the event "Abe knows that it is raining" entails "it is raining". The question is whether the event "Abe knows that it is raining" is self-evident for Abe and Beth. Because there seem to be ways for Abe to know that it is raining *without* Beth knowing that Abe knows that it is raining, the event "Abe knows that it is raining" is not self-evident (*mutatis mutandis* for a case with the event "Beth knows that it is raining" serving as the proposed self-evident event). It will also not work to take the self-evident event to be "Abe and Beth know that it is raining", for Abe may know that it is raining but not know that Beth knows that it is raining (because knowing a conjunction requires knowing the conjuncts).

In the end, finitary models seem to be no help supporting the possibility of common knowledge for realistic and non-ideal agents.[24] The above Abe and Beth case provides the best possible hope of common knowledge—if there ever is a case of common knowledge, this would be it. But in that case there are no plausible

[23]Because there are no windows in the basement, we can conclude that Beth doesn't know that it is raining. If Beth knew that Abe knew it was raining, it would be straightforward for her to infer that it is raining. So we should think that Beth doesn't know that Abe knows it is raining.

[24]Vanderschraaf and Sillari (2009) seem to suggest that finitary models may help secure the existence of common knowledge for real agents, but they do not give worked out arguments.

self-evident events that feature the appropriate entailment relation, so, by Finitary Common Knowledge, there is no common knowledge. Because the Abe and Beth case provides the best hope for common knowledge and there is none, common knowledge is not possible.

What is the allure of finitary models? What they seem to accomplish is, in effect, a distillation of the complexity of an infinite sequence of higher-order knowledge to a finite property. But the complexity isn't removed, it is only focused and repackaged. It sounds epistemically demanding to satisfy the infinite common knowledge hierarchy. It initially sounds less demanding to satisfy a self-evident event (perhaps because there is no infinite hierarchy). But the only self-evident events which feature the appropriate entailment relations require that agents have deep and sophisticated knowledge of other agents' knowledge. These agents must know that every other agent knows some event, *whenever* that event is known. Real, non-ideal agents cannot meet this demand. There are no self-evident events in standard cases[25] of common knowledge, so common knowledge is unattainable. From these remarks, I suggest that premise (3) is plausible.

2.3.4 Ensemble

Returning to the paradox, I've argued that each of the following are plausible:

> (1) We often rationally solve coordination problems.
>
> (2.2) If one does not know that other player knows some part of the common knowledge sequence for p then it is not rational to coordinate with solution R based on p.[26]

[25] On page 41, I framed the argument as against there being "non-trivial" common knowledge. One might argue that tautologies or logical truisms *are* common knowledge. Such a result follows from the set-theoretical model of knowledge as I've presented it. In Chapter 3 I discuss the issue of logical omniscience and whether we should think that agents know all tautologies.

[26] I previously suggested that (2.2) is equivalent to the original premise (2).

(3) Common knowledge is unattainable.

Because premises (1), (2.2), and (3) are jointly inconsistent, one of them must be false. In what is to follow I will diagnose the source of the problem and argue for a resolution to the paradox.

2.4 A Context Dependent Solution

I will argue that the best resolution to the paradox is to reject premise (2.2), and, correspondingly, premise (2). I take it that premise (1) is the least controversial of the three. David Lewis' pioneering work on convention is predicated on our rational coordination—his project was to explain and analyze how such rational behavior is possible. Moreover, coordination appears commonplace when we have the right lens to look for it. As social beings, we require cohesion, organization, and predictability, so our interests at least coincide in these dimensions.

Next, I attempted to argue forcefully for premise (3), beyond its mere plausibility, because I think it is true. When the subject is real, non-ideal agents like ourselves, common knowledge is not attainable. Why, then, has common knowledge played such an important role in philosophy and the social sciences? The best explanation is that technical concepts like common knowledge have a kind of theoretical usefulness because they idealize away the messiness of our cognitive limits and our actual practice of knowledge attribution. Common knowledge is a logician's concept, it can be formally described and analyzed. But, as Williamson (2000) suggests, "common knowledge would therefore be a convenient idealization, like a frictionless plane".[27] The analogy is that frictionless planes do not exist, but physical models of them are nevertheless essential to a working engineer, because of their pragmatic utility. Perhaps some philosophical insight can be gained from investigating common knowledge, even though it does not exist for non-ideal agents.

[27] Williamson (2000), p. 122.

2.4.1 Desiderata

A proper resolution to the paradox calls for more than showing which premise is false. After all, premise (2.2) was plausible. What might replace premise (2.2)?

The above discussion of the paradox makes two issues clear. First, rational coordination requires *some* knowledge. From the Coordinated Attack problem, it would not be rational for general A and general B to attack at dawn if no messages were sent. Likewise, if I wished to meet a philosopher at a coffee shop to discuss a paper, sent her an email but never heard back, in most cases it would not be rational for me to go to the coffee shop. Second, I think a proper resolution to the paradox should be able to capture the intuition we have about the Coordinated Attack problem. I take it that the intuitive response to Coordinated Attack is that the two generals will never be able to rationally coordinate an attack.

I will take these observations as desiderata for a proper resolution to the paradox. I turn to my solution next.

2.4.2 Context

Denying premise (2.2) and premise (2) yields the following "truncated" premise:

> (T) Rationally solving coordination problems does not require common knowledge.

In light of the two desiderata, (T) is insufficient, so it needs to be amended.[28]

Here are two ways to amend (T) that *do not* meet our requirements. First, suppose one were to propose that rational coordination always required knowledge up to the n^{th} member of the common knowledge sequence. For example, let n be 6 when there are two members in the group $\mathcal{G}$. This proposal won't explain our intuition on the Coordinated Attack problem. Second, suppose one were to

[28](T) is consistent with requiring no knowledge for coordination.

propose that it is simply vague as to how much knowledge is needed for rational coordination. This proposal doesn't clearly explain our intuition on the Coordinated Attack problem. But it also doesn't fit our practice of admonishment for lack of knowledge. If general A were to attack at dawn without receiving a reply from general B, we could reprimand him because he acted irrationally. This does not seem to be a case of vagueness.

What's needed to remedy (T) is a kind of flexibility. Sometimes, it seems, we need to know a lot to rationally coordinate, but other times we need to know less. Consider the familiar exchange of emails when attempting to find a meeting place to discuss a paper. Suppose I wish to meet with April to talk about our epistemology paper (further, suppose it is imperative that we meet because we need to resolve a draft before our presentation Saturday morning). For sake of concreteness, suppose that we're at a philosophy conference in New York City, in an unfamiliar area, and we're staying at different hotels.

I send April an email and ask her to meet me Friday at nine in the evening at the campus coffee shop. April sends a reply email with "great, see you there". In most cases this is all the information we would need to coordinate. Probably, we would both plan to meet at the coffee shop at nine, and we would be rational to do so. But what's interesting about this ordinary case is that it is nearly identical in structure to Coordinated Attack. We both know email is not perfectly reliable. Just as there was a chance that the messenger was captured in Coordinated Attack, there is a chance that each email is not delivered. Or, I might not have checked my email after April sent her reply. Yet in normal circumstances such considerations do not seem to matter—two emails are enough to rationally coordinate.

Now, suppose that it is extremely important that April and I meet to discuss our paper. Suppose we know that if we don't meet Friday evening before our presentation, we won't have a reply to a potentially devastating counterexample, and we'll completely embarrass ourselves and ruin our careers. In this case my intuition

is that two emails are not enough. I might think: "April didn't get a reply from me to her reply, maybe she thinks I've changed my mind". She might think: "I don't know if he received my reply or maybe he didn't check his email". What to do in such a case? We might send another email, a text message, or make a phone call. But, however it is achieved, given the importance of our meeting it seems rational for us to try to resolve some of our uncertainty about what we know. Note that in light of the previous section, all of these efforts must fall short of establishing common knowledge.

In line with these remarks, I suggest the best way to remedy (T) is to add a contextual element. The insight from epistemic contextualism is that context renders some possibilities of error salient or relevant. In a generic form, epistemic contextualism claims that context determines the amount of justification needed for knowledge.[29] In ordinary contexts, such as an evening with friends, the justificatory bar for knowledge is set rather low. In these contexts it is true for me to say that I know I have hands (or, more realistically for a social setting, that I know the coffee is still hot, say). But in more demanding contexts, such as a philosophy classroom, the justificatory bar for knowledge is set very high. In these contexts, epistemic possibilities like evil geniuses and brains in vats loom, and it may not be true to say that I know I have hands.

These general observations about context and justification can be repurposed for a premise about rational action. The basic idea is that in ordinary contexts we need less knowledge to rationally coordinate. Yet, when the possibility of error becomes salient, or the stakes for coordination are high, we enter a more demanding context and need more knowledge to rationally coordinate. Such mechanics help provide

[29]Epistemic contextualism is typically viewed as a semantic thesis about the truth conditions for a sentence or utterance of the form "S knows that p". Different versions of epistemic contextualism differ on the details, but they all have in common that more demanding contexts require a stronger epistemic position for knowledge.

the requisite flexibility needed to meet the desiderata.

Consider the following replacement for premise (2.2):

> (C) Rationally solving coordination problems requires knowledge up to the n^{th} member from the common knowledge sequence, where n is determined by context c.

Similar in spirit to epistemic contextualism, less demanding contexts make fewer possibilities of error salient for the group $\mathcal{G}$. In ordinary contexts, we fail to consider whether an email server is down or whether there is a power outage or whether the recipient's computer has crashed. And in these ordinary contexts, two emails are sufficient for rational coordination. Though I wish to remain neutral on a particular version of a context dependent premise, either raising the stakes or making some possibilities salient seem to result in a more demanding context and correspondingly, by (C), require more knowledge for coordination. When failing to coordinate is very costly, it seems rational to either send the third email, or make a phone call. When it becomes salient that the email server might be down, it seems rational to call.

Next, I will show that (C) can account for the desiderata. First, as given, (C) always requires *some* knowledge for rational coordination because n is always positive.[30] Second, (C) can plausibly explain our intuition in the Coordinated Attack problem. Typically, the full import of Coordinated Attack is not immediately obvious—it is gleaned in a stepwise fashion. It is natural to first think, "general A need only send one message". But it then becomes clear that A wouldn't know whether B received the message. So, thought turns to what would change if B sent a reply and further iterations of this process.

What seems to be happening is that each successive worry about the delivery of the next message makes salient both the possibility of the messenger being captured *and* the high stakes of error (the defeat of the army division). Context dependent

[30]Note that there is no zeroth member in the common knowledge sequence.

solutions are well suited to explain our reasoning in this case. Worry about the messenger being captured results in a more demanding context, which is tantamount to, by (C), requiring that further members of the common knowledge sequence be known for coordination. The next natural thought is to send further messages between A and B, which results in more knowledge from the common knowledge sequence. But then we worry about whether these messages will be received, and the process continues.

2.5 Knowledge and Action

My proposed solution can help inform a recent debate in epistemology about the connection between knowledge and action. Since Williamson (2000) there has been renewed interest in the relationship between accounts of knowledge and rational action. For instance, Williamson has argued that knowledge, and not mere justified belief, features ineliminably in explanations of action.[31] Hawthorne (2004) and Stanley (2005) have argued that the practical stakes of acting can influence knowledge attributions. Fantl and McGrath (2009), building on previous work, have defended various principles connecting knowledge and justification for action.

I've argued that the best resolution to the common knowledge paradox is to reject the second premise. So, rational solutions to coordination problems do not require common knowledge. From the definition of common knowledge this means that it is possible to have a rational solution to a coordination problem (act rationally) even though some member of the common knowledge sequence is false—i.e., some person doesn't know something about what they or the other member of the group knows. If this is correct, then this provides a counterexample to a much discussed principle proposed by John Hawthorne and Jason Stanley.

Hawthorne and Stanley (2008) defend the following principle:

[31] See Williamson (2000), p. 60–64.

> *Action-Knowledge*: treat the proposition that p as a reason for acting only if you know that p.[32]

One way to defend Action-Knowledge is by referencing our practice to admonish a person's acting by citing their lack of knowledge (just as I did, above, when I gave the ranking officer example). This is what Hawthorne and Stanley say:

> The Action-Knowledge principle makes immediate sense of our use of 'know' to criticize the actions of others. When someone acts on a belief that does not amount to knowledge, she violates the norm, and hence is subject to criticism. That is why we use epistemic vocabulary in criticizing the actions of others.[33]

Again, supposing that general A attacks at dawn without receiving a message from general B, it seems natural and appropriate for a ranking officer to criticize A by citing his lack of knowledge: "you didn't know that B was going to attack at dawn, so you shouldn't have attacked".

Before I examine Action-Knowledge, I need to say something about the kind of norm expressed by the principle. Hawthorne and Stanley say:

> Our principle concerns what is *appropriate* to treat as a reason for action, rather than what one *ought* to treat as a reason for action. . . The principle is therefore a claim about what is permissible to treat as reasons for action in a given choice situation.[34]

Further, to be clear, Hawthorne and Stanley need to say that the principle concerns what is *rationally* appropriate to treat as a reason for action.

[32]Hawthorne and Stanley (2008), p. 577.
[33]Hawthorne and Stanley (2008), p. 577.
[34]Hawthorne and Stanley (2008), p. 578, original emphasis.

Action-Knowledge cannot be correct if my solution to the common knowledge paradox is correct. This is because Action-Knowledge cannot make sense of the difference in our practice to criticize the generals in Coordinated Attack (supposing only two messages are sent) but not to criticize when only two emails are sent in low stakes contexts. Recall the structural similarity between Coordinated Attack and typical cases of coordination by email that I mentioned in section 2.4.2. The only relevant difference in these two cases is the magnitude of the cost and benefit of coordinating. It is clear that we *can* criticize the generals for acting if only two messages are sent but it is also clear that we *cannot* criticize April for going to the coffee shop when only two emails are sent. Again, if the stakes are low and I send April an email asking her to meet me at the coffee shop at noon to discuss a paper (supposing that she replies with the message "great, see you then"), it would be rationally *inappropriate* to criticize her by pointing out that she didn't know whether I received her reply (moreover, it is rational for her to go to the coffee shop). Action-Knowledge cannot make sense of our use of 'know' in this case and this undermines a central reason we may have had to accept the principle.

Further, it is revealing that we cannot criticize April for showing up at the coffee shop after two emails are sent. Again, it would be rationally *inappropriate* for me to admonish April at the coffee shop with the following: "you shouldn't have gone to the coffee shop because you didn't know whether I received your reply". Note that this is *not* the inappropriateness of etiquette but, instead, that of having reason for acting. Why can't we criticize April in this case? I submit that the best explanation for our not being able to criticize April is that it was rationally permissible for her to act on the proposition "he received my reply", even though she didn't know that I received her reply. Certainly, we should judge that she *expected* me to be at the coffee shop (otherwise, why would she have gone to the coffee shop). The best explanation for her expectation was that she believed that I received her reply (otherwise, why would she expect me to be at the coffee shop). Given that we judge April to be

rational for going to the coffee shop, her deliberation seems beyond reproach and, hence, rationally permitted.

If my context dependent solution to the paradox is correct, then there will be cases where some element of the common knowledge sequence is not known, yet it was rationally permissible for the members of the group to treat that proposition as a reason for coordinating. This serves to break the connection between knowledge and action that Hawthorne and Stanley seek to defend. As it turns out, cases of group coordination show that one may rationally act *without* knowledge of propositions directly relevant to the outcome. From the paradox, if we needed to know *everything* relevant to a coordinating outcome, we would need to have common knowledge. But we don't have common knowledge so we can rationally act despite not knowing.

Hawthorne and Stanley may respond by denying my solution to the common knowledge paradox. But then they would be saddled with the paradox once again—if premise (2) is true then either premise (1) or premise (3) is false. A proper defense of Hawthorne and Stanley's position, then, must either argue that we can possess common knowledge or that we don't rationally solve coordination problems. More is at stake between knowledge and action then has been previously acknowledged.

2.6 Lewis on Common Knowledge

As I previously mentioned, David Lewis (1969) famously argued that conventions in language and action must be common knowledge. Yet Lewis was keenly aware that human agents face a variety of cognitive limitations such as finite memory and processing power. How might Lewis respond to the common knowledge paradox?

Lewis (1969) is commonly cited as being the first to articulate the concept of common knowledge.[35] Many texts in game theory, especially, give Lewis credit

[35]However, as Cubitt and Sugden (2003) point out, Robert Nozick (2001), p. 375, fn. 60, attributes to himself the first formal characterization of common knowledge in his 1963 doctoral dissertation.

for the idea that common knowledge is to be understood as an infinite hierarchy of interpersonal higher-order knowledge; that is, the conjunction of the common knowledge sequence I gave in section 2.3.2. However, it is perhaps unsurprising that Lewis' view is actually more complicated than is typically acknowledged. While I do not intend to engage in deep exegesis of Lewis' *Convention*, I want to highlight several features of Lewis' view and relate them to the common knowledge paradox.

Even though Lewis is routinely credited with being the first to formalize a notion of common knowledge, he is not ostensibly discussing knowledge. He is clear in making a distinction between *belief* and *having reason to believe*, and he is clear that his analysis concerns what agents have reason to believe. Roughly, reasons to believe appear to be something like *potential* beliefs of agents, perhaps what an agent would believe were she to have more time and computational resources. Cubitt and Sugden (2003) interpret "reason to believe" in the following way: "to say that some individual i has reason to believe some proposition x is to say that x is true within some some logic of reasoning that is *endorsed* by (that is, accepted as a normative standard by) person i".[36] Correspondingly, they suggest that Lewis' account isn't really about knowledge, but about reasoning or justified or warranted belief. Lewis later remarked that his choice of terminology in 'common knowledge' was imperfect: "that term [common knowledge] was unfortunate, since there is no assurance that it will be knowledge, or even that it will be true".[37] That is, Lewis acknowledges that real human agents may have reason to believe a proposition, even though the proposition is false.

Before I continue to discuss Lewis' account of common knowledge, it is worthwhile to consider the implications of the distinction between belief and having reason to believe. If Lewis was discussing some form of mere potential belief (what we might otherwise reasonably believe) as underwriting rational coordination,

[36]Original emphasis. See Cubitt and Sugden (2003), p. 184.

[37]See Lewis (1978), p. 44, n. 13.

the complexity of an infinite hierarchy of interpersonal epistemic claims (I believe that you believe that I believe...) does not seem to be overly problematic. As Vanderschraaf and Sillari (2009) interpret Lewis, the infinite hierarchy of interpersonal knowledge or belief claims really amount to "chains of logical consequences", not actual steps of reasoning that the agents must go through and consider. Returning to the common knowledge paradox I've presented, we might think that Lewis would simply deny premise (2.2) (and, correspondingly, premise (2)). Perhaps, in Lewis' view, rational coordination does not require *actual* interpersonal knowledge as given in premise (2.2), but only that agents are so situated that either they *have* the appropriate infinite hierarchy of reasons (but might not be able to appreciate them) or that they *could* complete every chain of logical consequences from the infinite hierarchy (but do not). Might either of these suggestions work to provide a replacement for premise (2.2)?

To consider these suggestions, it is helpful to return to the familiar example of coordination by way of email. Suppose A wishes to meet B at the coffee shop at noon, so A sends B an email with the message "let's meet at the coffee shop at noon". Now, as things currently stand, A does not know whether B has received the email. I think it is intuitively obvious that A would not be rational to go to the coffee shop at this stage. But suppose that A and B were so situated to *have* the infinite hierarchy of interpersonal reasons, even though they cannot appreciate or access them. I think it is likewise obvious that it is *not* rational for B to go to the coffee shop (hence it is not rational for A and B to coordinate). So, merely having reasons is not enough for rational coordination. Instead, suppose that A and B were so situated such that they could complete every chain of logical consequences from the infinite hierarchy of interpersonal epistemic reasons but do not. It still seems to me that it is not rational for A and B to coordinate.

The above considerations seem to suggest that it is *actual* knowledge that matters for rational coordination. Merely having unappreciated reasons or the ability to

infer various epistemic claims does not make one immune from criticism. Suppose, for instance, that general A from Coordinated Attack had the common knowledge sequence available concerning some relevant proposition p (A knows p, B knows p, A knows that B knows p, ...). That is, suppose A could access this information if he were to approach a nearby oracle, but A does not. Suppose A attacks at dawn. It then seems that A acted irrationally. If a ranking officer discovered that A acted without consulting the oracle, it seems she would have grounds to admonish A. The ranking officer is able to criticize A by citing that A didn't know whether B knew to attack at dawn.

While it might be judged that Lewis (1969) would deny premise (2.2) and premise (2) from the common knowledge paradox, it is unclear whether his work helps provide a proper resolution. Of course, this was not the purpose of *Convention*. In his 1969 book, Lewis set out to *explain* how certain regularities of action were possible. For example, it is a social convention that Americans drive on the right side of the street. Lewis' insight was to explain such behavior by invoking, roughly, a notion of informational stability in the coordinating agents. Lewis realized that if agents *would* reason along similar paths (even if they never *actually* completed such reasoning), their behavior could be predictable enough to be relied upon. Roughly, Lewis understood a convention as arising when and only when it is common knowledge that there is some regularity R of behavior such that:

(i) everyone conforms to R

(ii) everyone expects everyone else to conform to R

(iii) everyone has similar enough preferences.[38]

So, the explanation for the regularity of behavior by Americans to drive on the right is that there is some, again, roughly, stable state of expectations among the

[38]This is not Lewis' actual analysis, but a simplified account. Lewis' account is complicated and a proper reconstruction is beyond the scope of this present work.

population. The thought is that *were* anyone to reason about their actions, their reasoning would support participating in the convention. And because everyone *would* reason in this way, their beliefs and actions are stable and reliable, so they support the behavioral regularity of participating in the convention. Plausibly, this works to *explain* how conventions could arise and persist.

The common knowledge paradox I've presented has a different aim. Premise (2) and premise (2.2) *seem* plausible. The thought is that if an agent doesn't know some member of the common knowledge sequence, then there is some amount of uncertainty about her actions. Returning to the above example of coordinating by email, if A never received a reply from B that she received the email, then it is irrational for A to coordinate. Note that in this case, there is no convention on which either A or B may rely. Matters might be entirely different if it was previously established that, say, A and B always go to the coffee shop at noon. Conventions provide one way to coordinate, but absent a convention, it seems that agents need some amount of *actual* knowledge to provide for the missing stability. My context dependent solution given in section 2.4.2 helps account for how much knowledge is needed.

2.6.1 Ernst on Coordination and Heuristics

In an unpublished paper, "Convention and Bounded Rationality", Zachary Ernst (2012) also considers the question of rational coordination for real, non-ideal agents. Similar to the coordination by email example, Ernst considers two agents, Mark and Bob, who wish to arrange a clandestine meeting. They have previously arranged a system in which Mark is to put a flag outside his window when he wishes to meet. When Bob sees the flag, they will meet at a particular parking garage at a prearranged time. Suppose, further, that there is a risk of showing up alone (they would draw unwanted attention to themselves), so neither Mark nor Bob wants to show up alone. What must Mark and Bob know to rationally coordinate?

Suppose Mark hoists his flag and Bob later sees this. Ought they coordinate? Ernst thinks "no". While Mark and Bob might have mutual knowledge that the flag is raised, neither Mark nor Bob knows that the other knows the flag is raised. Ernst then suggests that Bob might signal to Mark that he saw the flag. Suppose that they also previously agreed that when Bob sees the flag he moves a flower pot on his window ledge. Suppose Bob moves the flower pot. Should he then go to the garage? Again, Ernst thinks "no". Bob doesn't know whether Mark saw the flower pot.[39]

So far, Ernst agrees with much of my assessment of the common knowledge paradox. Indeed, his flag case is structurally similar to Coordinated Attack and the above case of coordination by email. Further, Ernst thinks that in real world instances of his flag case, Mark and Bob would manage to rationally coordinate (perhaps, just as we rationally coordinate by way of email). But Ernst's end assessment is different from mine. He says,

> Surely... any rational individual will coordinate in such a situation while having only one (or at most, two) levels of interactive knowledge... we successfully act without making such strong demands for reassurances. However, the reason why we do not make "unreasonable" demands for an increasing number of reassurances is because we are not typically engaged in a process of reasoning in accordance with these "rational" requirements at all. Rather, we approach such situations in the real world through a set of simple heuristics that we take as sufficient for acting...[40]

[39]In the above, Ernst is giving the reasoning in accordance with what he calls the "standard notion of rationality".

[40]See pp. 13–15 of the December 2009 version of Ernst's unpublished draft "Convention and Bounded Rationality". Ernst uses the term 'reassurances' to mean additional requests for interpersonal higher-order knowledge.

First, it is clear that Ernst would also deny premise (2.2) and premise (2) from the common knowledge paradox. From the above quote, he thinks that we can rationally coordinate with at most two levels of interactive knowledge. Second, he seems to think that we rationally coordinate not by explicitly reasoning about the knowledge requirements for rational coordination but by simple heuristics. I will explore each of these responses below.

I think Ernst is wrong in claiming that real world coordination problems similar to the flag case or the email case or Coordinated Attack require at most two levels of interactive knowledge. Indeed, Coordinated Attack provides an excellent counterexample. It is intuitively plausible that the generals in Coordinated Attack require more than two levels of interactive knowledge to rationally coordinate. This is because the stakes are incredibly high—if one general acts without the other, his entire division will be killed. Even if several messages are delivered between them, it still seems irrational for them to act—each general knows that there is some chance that the last message will not be successfully delivered. Now, if the generals *were* able to meet face to face (such that general A could say directly to general B "we attack at dawn"), it then seems rational for them to coordinate. Hence, because it is now rational to coordinate (given the face to face communication), we can infer that the generals now have *more* than two levels of interactive knowledge. And because the additional knowledge is the only relevant difference in the two situations, it is clear that rational coordination in the high stakes version of Coordinated Attack requires more than two levels of interactive knowledge.

As further illustration, Suppose that Mark and Bob face a different coordination problem. Suppose that if either Mark or Bob show up alone at the parking garage they will be killed. In such a case it seems entirely reasonable for Mark and Bob to make "strong demands for reassurances". The point is that as the stakes for unsuccessful coordination grow, it *is* reasonable to demand more reassurances in the form of more interpersonal knowledge. This is the insight captured in my solution

to the common knowledge paradox, given in premise (C) in section 2.4.2. And though we might not typically face such grave coordination problems as overthrowing governments or leading army divisions, our concept of rational coordination does extend to cover these cases.

Next, in light of these observations, it is unclear how helpful Ernst's suggestion that we rely on simple heuristics is for resolving the common knowledge paradox.[41] Ernst may be correct in thinking that we typically do not *explicitly* reason about what we and others know about our and others' knowledge in standard cases of coordination, but this does not mean that we cannot reason explicitly about higher-order interpersonal knowledge nor that we should not. In fact, Coordinated Attack shows that when the stakes are high it becomes obvious and relevant to explicitly inquire about higher-order interpersonal knowledge. Even if we typically use heuristics as a guide in our coordination, Coordinated Attack provides a case where we seem to override any simple heuristic strategy. Further, it seems like this is the rationally appropriate move to make.

It is unclear to me to what extent my context dependent solution (C) is compatible with a heuristics-based view of rational coordination. Perhaps, in typical cases, our "coordinated action heuristic" is sensitive to features such as salience of error or high cost of unsuccessful coordination. But, often, theoretical uses of heuristic strategies are just black boxes. And, as always, the details will matter. Specifically, we need to know when it is rational to override our use of heuristics. Coordinated Attack seems to suggest that we sometimes explicitly notice that we need more higher-order interpersonal knowledge to coordinate. This problem mirrors the broader problem of rationality, concerning the interrelation between high-level rationality (explicit reasoning) and low-level rationality (heuristics).

[41] Of course, Ernst is not addressing the common knowledge paradox as I've given it. But he is addressing the issue of how real agents rationally solve coordination problems.

2.7 Conclusion

I have argued that a proper resolution to the common knowledge paradox is to reject premise (2.2) (and thereby reject premise (2)). As a replacement, my context dependent premise (C) requires *some* knowledge of the common knowledge sequence, but not knowledge of the entire sequence to make coordination rational. This solution also has the virtue that it can explain our intuition in the Coordinated Attack problem.

Further, I've shown how my solution connects with the recent discussion of knowledge and action. Our rational solutions to coordination problems show that we can treat some propositions as a reason for acting even when we don't know the proposition.

CHAPTER 3

THE CLOSURE OF KNOWLEDGE, OMNISCIENCE, AND AWARENESS

3.1 Introduction

Deduction seems like a way of extending our knowledge. If I know that q follows from p, and I know that p is true, then it seems like I can come to know that q. Supposedly, mathematical discovery fits a similar pattern. Deduction also seems to help us understand what others know, what their view consists in. When S knows p, and S knows that p entails q, it often seems reasonable that we should say that S knows q. For example, if Sarah knows that Bill is either in the kitchen or in the bedroom, and she knows that he's not in the bedroom, then it seems reasonable to judge that Sarah knows that Bill is in the kitchen.

To say that deduction can extend our knowledge is to (plausibly) give credence to some kind of closure principle. To begin, consider the following much discussed principle:

(A) If S knows p, and p entails q, then S knows that q.

If (A) were true then our knowledge would extend far beyond what we typically acknowledge. For instance, suppose Sarah knows the Peano axioms for arithmetic (suppose she knows these by testimony from her mathematics professor). By (A) Sarah would then know many complicated theorems about natural numbers that follow from the Peano axioms. Actually, Sarah would know *every* theorem that follows from the axioms, including theorems that haven't been formally proven yet (or even considered) and theorems that cannot be proven in a human lifetime. This result is highly counterintuitive. Surely, Sarah doesn't know everything that follows from the Peano axioms. Hence, (A) is likely false.

Yet, contrary to the above, there are versions of (A) that *are* highly plausible. Consider the following:

(A′) If S knows that p, and p entails p, then S knows that p.

Call (A′) the principle that knowledge is closed under "self entailment". Though this principle is not very interesting (it is trivial, perhaps), it shows that instances of (A) are acceptable and that there are *some* true closure principles.[1]

If we assume that deduction can extend our knowledge in some meaningful way, a resolution to the issue of the closure of knowledge begins with looking for informative principles situated between the extremes of (A), which is too demanding (we don't know every deductive consequence of our knowledge), and (A′), which is not demanding enough (deduction does seem to *extend* our knowledge). The hunt is for a fertile middle territory. However, I want to flag at the outset some likely jagged and rough features of this middle terrain. As it is with many geographical objects, the base and summit offer predictable landscapes, while the middle offers peril. The story of deductive closure will prove similar. At the extremes, perfectly ideal agents might best be thought to know all deductive consequences of their knowledge, while perfectly unintelligent, simpleminded, and dull-witted agents might best be thought to know none of the deductive consequences of their knowledge. Human agents, somewhere between the extremes, provide difficult ground.

I will also investigate what is called "the problem of logical omniscience", and discuss its relation to issues concerning the closure of knowledge. Many epistemic

[1]Below I discuss closure principles more generally. One might argue whether (A′) actually establishes that there are true closure principles. Certainly, the self entailment relation can be true or false for any proposition p, just as the entailment relation can be true or false for any pair of propositions p and q. In this regard, knowledge *is* closed under self entailment and so knowledge *is* closed under some relation. However, I agree that (A′) does not go any way toward an understanding of the platitude that deduction is a way of *extending* our knowledge. I turn to this issue in the next section.

logics have, as a consequence of their semantics, properties similar to (A). For instance, in a standard S5 epistemic logic, if a proposition p is known, and p entails q, then q is known as well. Many have argued that such properties show these epistemic logics to be *inappropriate* models of human knowers. The *problem* of logical omniscience is to find modifications to these logics that capture realistic features of human cognizers.

In the end, I will argue that the central issue behind the closure of knowledge and the problem of logical omniscience concerns the amount of idealization in our models and our theorizing. This conclusion is apparent after the parallels between closure and logical omniscience are made clear. I next consider several ways to model realistic features of human cognition and show how they impact whether knowledge is closed.

3.2 Closure

Begin with the platitude that deduction is a way of extending our knowledge. A mathematician knows p, gives a proof of q from p, and thereby comes to know q. This seems unproblematic. Yet, it *is* a problem to clearly articulate the platitude. As I mentioned in the introduction, one may be tempted by the following:

(A) If S knows p, and p entails q, then S knows that q.

Clearly, (A) is insufficient. We know many things, consequences of which we've never considered or entertained. Again, suppose Sarah knows some mathematical axiom p, and that p entails q. By (A), Sarah knows q. But suppose that Sarah has never considered q and that q is so complicated it would take days to parse, and thousands of steps to prove. Intuitively, Sarah does not know q even though q follows from what she knows.

If (A) is false, then the lesson is that knowledge is not closed under entailment alone.[2] The term 'closure' is borrowed from mathematics. In mathematics, a set S is closed under operator $\mathcal{O}$ if and only if for any element $s \in S$, the operator $\mathcal{O}$ returns some element $s' \in S$ on the input s.[3] As an example, consider the set of natural numbers $\mathbb{N}$ (the set of positive counting numbers, or positive integers). The set $\mathbb{N}$ is closed under the addition function "+". Take any two $a, b \in \mathbb{N}$. Then $a + b = c$, because "+" is well-defined. And $c \in \mathbb{N}$, because $c \geq a+b$. Simply, $\mathbb{N}$ is closed under the addition function "+" because adding natural numbers always yields another natural number. To contrast, the set $\mathbb{N}$ is *not* closed under subtraction. For instance, 2 and 9 are natural numbers, but $2 - 9 = -7$, and $-7 \notin \mathbb{N}$. $\mathbb{N}$ is not closed under subtraction because subtracting two natural numbers does not always yield another natural number.

Questions about the closure of knowledge have similarities to these mathematical questions. Revisiting Sarah, consider the set of propositions known by Sarah (call this set "$\mathbb{K}$"). We may ask whether $\mathbb{K}$ is closed under various "epistemic operations". In assessing (A), I argued that, for Sarah, $\mathbb{K}$ is not closed under entailment. That is, there are propositions entailed by what Sarah knows that are not, themselves, known by Sarah, and, hence, not members of $\mathbb{K}$.

Entailment is one example of an "epistemic operation". Consider another epistemic operation: known entailment. Modifying (A), we would have:

(B) If S knows p, and S knows that p entails q, then S knows that q.

This principle escapes the worries posed for (A) because the "epistemic operation" of known entailment effectively strengthens S's epistemic position—in (B), S *knows* that p entails q. To contrast, S's epistemic position in (A) is weaker with respect to the entailment because we do not assume that S knows that p entails q.

[2]Below I will consider variants of (A).

[3]The definition naturally extends for higher arity relations.

While (B) is intuitive, there are still problems with its formulation. One problem is that (B) does not ensure that S notices the connection between p and what p entails. Consider an example. Sarah comes to know axiom p from testimony by her college number theory professor. Later in the semester Sarah notices that p entails q (by a complicated proof, say). But, suppose at the time of her proof, she doesn't notice that her proof is related to what she learned from her professor (that p). That is, she mistakenly thinks that her knowledge that p and her knowledge that p entails q are unrelated pieces of information. It then seems reasonable to judge Sarah as not knowing q.[4]

What comes next should not be surprising. The literature on epistemic closure has the familiar feel of the literature on the analysis of knowledge. Analysis A is offered, only to be met with counterexample C, followed by modified analysis $A^\star$ and counterexample $C^\star$, and so forth. At this stage it is important to take stock of the goal and the motivation.

The motivation behind my investigation of closure principles is to understand and clarify the role that deduction seems to play in expanding our knowledge. The goal, however, is less clear. Some have suggested that the search for true closure principles is the search for an exceptionless principle.[5] If this is the goal then the dialectic will surely take the structure of the analysis of knowledge: proposed analysis and potential counterexample. But I will suggest that there is another goal to aim for, one that can illuminate the connection between deduction and knowledge. Such a goal becomes clear when we turn to the problem of logical omniscience.

[4]In section 3.3.1 below I give a similar example of a student who is not able to see the connection between his proof, his assumptions, and what he is trying to prove.

[5]See Kvanvig (2006).

3.3 The Problem of Logical Omniscience

Early development of epistemic logics, logics of knowledge, focused on the modal character of knowledge. Hintikka (1962) found a similarity between knowledge and necessity. Just as a modal claim's truth (such as "$\Box p$") depends on p's truth in other possible worlds (or states), so too for knowledge. One may flesh out the idea in various ways, but the core of the idea is that S knows that p if and only if p is true in all epistemically possible worlds (or states) "accessible" to S. This core idea shares conceptual territory with relevant alternatives theories of knowledge which came later: for example, Dretske (1970), Goldman (1976), and Stine (1976). If we interpret "accessible worlds" (or states) as those an agent cannot appropriately rule out or exclude, then an agent S knows a proposition p if and only if S can exclude all non-p worlds.

The contemporary way of formalizing this notion can be expressed in a Kripke structure with the following:

$$(M, s) \models K_i\varphi \text{ if and only if } (M, t) \models \varphi \text{ for all } t \text{ such that } (s, t) \in \mathcal{K}_i.$$

In words, this says that an agent i knows φ at state s if and only if φ is true at every accessible state t. Roughly, the idea is that i knows φ when and only when every way the world could be (to the agent i) is a world where φ is the case.[6]

Yet, such a formalization has several consequences. Call the first property "closure under logical implication" (note the similarity to (A) as previously given):

(CLI) If S knows φ, and if φ logically implies ψ, then S knows ψ.

Checking that (CLI) follows from the above formalization is straightforward. Suppose that S knows that φ and also that φ logically implies ψ. Because φ logically

[6] I previously introduced Kripke structures and the above characterization of knowledge in Chapter 1.

implies ψ, then ψ is true in any model where φ is true. Consider some t such that $(s,t) \in \mathcal{K}_i$. It must be the case that ψ is true at t. Since t was arbitrary, S knows ψ.

The next property standard in most epistemic logics is called "closure under logical equivalence":

> (CLE) If S knows φ, and if φ and ψ are logically equivalent, then S knows ψ.

To see that (CLE) follows, suppose that S knows φ and that φ and ψ are logically equivalent. By logical equivalence, φ and ψ are true in the same worlds. So if S knows φ, φ is true in all worlds epistemically possible for S. So ψ is then true in all worlds epistemically possible for S. Hence, S knows ψ.

The so called "problem of logical omniscience" is reconciling this formal framework for knowledge with its unrealistic consequences, such as (CLI) and (CLE), when the model is intended to capture and characterize, broadly, human knowledge. Fagin et al. (1995) remark, "people are simply not logically omniscient; a person can know a set of facts without knowing all of the logical consequences of this set of facts... One obvious source is lack of computational power...".[7] As an analogy, consider the game of chess. The rules for chess are involved, but simple enough for many young children to grasp them. The rules can be thought of as axioms. Knowing the rules allows one to progress the game by further application of the rules. For instance, the rules specify the initial configuration of the board, that white moves first, and that d4 is a permissible first move. Clearly, however, white (when human and unaided by computers) cannot know whether her complete strategy is a winning strategy *before* play begins, because such a proposition is too computationally demanding.[8] At best, white only sees the consequences of her position for several moves ahead (i.e., white does not know whether some future

[7] Fagin et al. (1995), p. 309–310.

[8] Note that chess *is*, in principle, solvable. With enough computational resources one could tell

move will be permissible). Further consequences are unavailable. Consider other observations about chess. There are systematic mistakes that some chess players may make, depending on the circumstances. One elementary mistake is to ignore an opponent's best response, and plan the next move as if some piece will not be captured (this is a low-grade form of wishful thinking). As well, there are occasions (due to anxiety or lack of time) when one cannot see that one is in a position to force checkmate.

In this limited respect, the analogy with chess and deduction is fitting. There is an intuitive sense in which proof in formal systems of logic is similar to a progression of a chess game.[9] We want to know whether some strategy wins, or whether some sequent has a proof. There are admissible moves in chess just as there are admissible new lines in a proof.[10]

In general, our cognitive limitations impose on us a kind of "cognitive haze" such that the first few moves ahead (or logical consequences) can be seen, but later moves are obscured. Andrew Wiles described his remarkable proof of Fermat's Last Theorem with a similar analogy. He described the beginnings of a complicated proof as walking around in a house with no lights, bumping into furniture. But as connections become clear, light dawns on the room and slowly, objects in the room come to be recognized.[11]

---

whether one could, at worst, force a draw with some strategy (a guarantee not to lose) at the beginning of play. Tic-tac-toe has recently been solved, so a sufficiently well-programed computer will never lose (interpreting a draw as "not losing").

[9]I do not mean here to endorse any metaphysical claims about mathematics or logic, such as formalism, in the philosophy of mathematics.

[10]Lewis Carroll, in *The Game of Logic*, develops a game using counters and diagrams to demonstrate the validity of various categorical syllogisms. See Carroll (1958).

[11]In the PBS NOVA episode "The Proof", Andrew Wiles says this: "Perhaps I could best describe my experience of doing mathematics in terms of entering a dark mansion. One goes into the first room, and it's dark, completely dark. One stumbles around bumping into the furniture, and gradually, you learn where each piece of furniture is, and finally, after six months or so, you

3.3.1 Goals

Return to the question of what we want for accounts of closure and logical omniscience. One may search for exceptionless principles, "close cousins" to (A) and (B), that comport with our concept of knowledge and capture what we mean when we say that deduction is a way of extending our knowledge.[12] But there is another goal in the vicinity, one that is, perhaps, more illuminating. Rather than searching for exceptionless principles, I aim to find a model that shows how our cognitive limitations impact various closure principles, one that is sensitive to the level of idealization present (I'll say more about this last part below).

Such a model should be able to explain the platitude that deduction is a way of extending our knowledge, but it would do so by showing how various cognitive limitations and systematic mistakes of reasoning rob us from knowing every logical consequence of our knowledge. In a broad sense, this is to recognize another platitude (or, at least, an important empirical observation): cognitive limitations prevent knowledge, so agents not capable of seeing the connection between what they know and what it implies do not know the implication of their knowledge. This empirical observation is easily learned by viewing performance in introductory logic courses. There are occasions when a student, finding himself on line n of a proof, has everything he needs to finish the proof but cannot "see", in some sense, how to put the pieces together. In the language of principle (B), this student knows p, knows that p entails q, but cannot see the connection between these two pieces of knowledge. Feeling a bit like Lewis Carroll's Achilles, one can say to the student[13]:

So you know p?

find the light switch. You turn it on, and suddenly, it's all illuminated. You can see exactly where you were" (PBS airdate, October 28, 1997).

[12]Kvanvig (2006) uses the expression "close cousin".

[13]The following exchange happened to me while I was an instructor of an introductory symbolic logic class.

"Yes".

And, you've proved that q follows from p?

"Yes".

So isn't it the case that q?

"...".

Reticence on the part of the student may come from nervousness, poor memory (perhaps impacted by nervousness), conceptual confusion about the "follows from" relation, or other cognitive factors. In some cases it may be that errors in syntactic processing do not allow the student to see that a rule applies. For instance, the rule modus ponens is actually a *schema*, representing an infinite number of particular instances. But the rule, itself, is syntactically represented by one instance (with the instance intended as a variable). This yields the possibility that a student knows the rule (in general) and knows the antecedent, but doesn't know that rule applies.

These considerations are important, especially when considering exceptionless closure principles. Timothy Williamson (2000) is credited with formulating a much discussed closure principle (which he calls "intuitive closure"):

(C) If you know p and competently deduce q from p, then you know q.[14]

The above case of the reticent student plausibly provides a counterexample to this principle. The student in question knows p, competently deduces q from p (in that he correctly and non-luckily applied the rules), but doesn't know q. There is surely room to debate what is meant by "competent deduction" and I'll return to this question in section 3.6.1.

I'm interested in the kinds of case where cognitive limitations systematically impact our failure to see the logical consequences of what we know. Surely such

[14]See Williamson (2000), p. 117, and Hawthorne (2005), p. 29.

features of cognition are relevant to the question of closure. But before I explore a model of cognitive limitations, I want to note a feature of the dialectic. When principles such as (C) are given, the motivating examples are often abstract and encourage idealization. For instance, the example I began with in section 3.1 was a case where Sarah knows that Bill is either in the kitchen or in the bedroom, and she also knows that Bill isn't in the bedroom. The natural and intuitive response is that Sarah knows that Bill is in the kitchen. Yet the case is silent on many relevant features. Is Sarah's cognition impaired? Perhaps Sarah suffers from the rare condition, call it "Sarah's syndrome", that causes her to not be able to process the left disjunct as an independent formula or sentence when she considers a disjunction. So, Sarah knows that Bill is either in the kitchen or the bedroom, and comes to know that Bill isn't in the bedroom. When Sarah considers the disjunction and the false right disjunct, she is unable to process the left disjunct and, hence, does not come to know that Bill is in the kitchen.

Of course, such a case is fanciful—it seems unlikely that there is a condition similar to "Sarah's syndrome". But even if there were such syndromes, the point is that in considering the original case we naturally idealize away such possibilities. We assume (perhaps tacitly) that Sarah is a competent reasoner, that she's not an infant, that she doesn't die right before making the inference that Bill is in the kitchen, and so forth.

The general problem of logical omniscience is that traditional formal models of knowledge are unrealistic and idealized. In essence, they abstract away the messiness and complexity of cognition. Next I will turn to Robert Stalnaker's (1999) account of the problem of logical omniscience and show how it can help inform the debate about the closure of knowledge. I will then give a model of knowledge that aims to capture cognitive limitations and show how closure and omniscience are impacted.

3.4 Stalnaker on The Problem of Logical Omniscience

Robert Stalnaker (1999) has given an assessment of the problem of logical omniscience.[15] He begins, rightly, with a discussion of idealization. One natural response to the problem of logical omniscience is to view formal models simply as an idealization in some unspecified way—this is to dissolve the problem. There is no problem of logical omniscience, the thought is, because formal systems are not intended to serve as models for real, non-ideal agents. But such a response misses the point. If the formal system is to say anything about knowledge, as we understand the concept, then there must be *some* problem of logical omniscience looming.[16] A theorist *could* say that her formal model has no connection to the ordinary sense of 'knows'—instead it is a model of 'knows*', a technical term. But this response is implausible. If there is no connection between knowledge, as we understand it, and knowledge*, then the project to model knowledge* seems unmotivated. A model of knowledge* would then be some uninterpreted formal system. Further, such a reply belies actual practice—constraints given in formal models of knowledge often appeal to rough yet widely agreed upon properties of knowledge such as factivity. Factivity, it is often cited, distinguishes knowledge from belief. That all formal models of knowledge provide for factivity speaks to there being some connection with the ordinary sense of 'knows'.

However, there are still issues concerning idealization. Stalnaker (1999) considers the response to the problem of logical omniscience that views knowledge in formal systems as "implicit knowledge". Begin with a rough characterization: call known propositions that are explicitly stored, perhaps written down and actually encoded in *mentalese*, "explicit knowledge", and call propositions an agent is committed

[15] See "The Problem of Logical Omniscience, I" and "The Problem of Logical Omniscience, II" in Stalnaker (1999).

[16] Part of Stalnaker's (1999) project, he claims, is to figure out just what the problem of logical omniscience amounts to.

to in virtue of having the explicit knowledge she does, "implicit knowledge". This rough characterization helps make sense of the following kind of case. Suppose Franz has read many ornithological works and knows that Harris's hawks are chocolate-brown and have chestnut shoulder patches (because he read this). We might also say, of Franz, that he knows that Harris's hawks in the wild don't wear tuxedos in winter. In the first case, it is natural to say that Franz knows explicitly that Harris's hawks are brown, because he has read and stored this information (it is, at least, a memory he can recall). But in the second case we should suspect that Franz has never considered or entertained whether Harris's hawks wear tuxedos, though such a proposition easily follows from what he knows about animals, birds, and formal evening clothes. Here we should say that Franz knows implicitly that Harris's hawks don't wear tuxedos in winter.

Suppose, now, one were to give a proposed solution to the problem of logical omniscience by claiming that that formal systems of knowledge are intended to capture a notion of implicit knowledge. Is this an adequate response to the problem? It is not, for several reasons. First, such a response does not make clear the platitude that deduction is a way of extending one's knowledge. By definition, implicit knowledge is closed under logical implication.[17] Yet, we don't reason with implicit knowledge—deduction on implicit knowledge doesn't extend what we implicitly know, because our implicit knowledge cannot be extended. Said differently, deduction doesn't seem like a way of coming to (implicitly) know new propositions because such implications are already implicitly known. Second, the notion of implicit knowledge seems too ideal to capture the second observation that agents are not always capable of seeing the connection between what they know and what it implies (and hence do not always know the implication of their knowledge).[18] The point of this observation motivates the problem of logical omniscience once again—that is, humans have

[17]'Implicit' and 'implication' both share a common etymology with the Latin 'implicitus'.

[18]Recall the previous example of the student not recognizing how to finish the proof.

inherent cognitive limitations and this impacts what they know. A move to discuss only implicit knowledge avoids the problem of logical omniscience by changing the subject.

However, there is a more serious objection to interpreting knowledge in formal logics as implicit knowledge. Stalnaker (1999), in several places, argues that the problem of logical omniscience is a problem of the accessibility or availability of information:

> The problem of logical omniscience, I am suggesting, is the problem of accessibility.[19] The manifest fact that we are not logically omniscient is a fact about our computational limitations—the fact that some of the information that is implicit in what we know or believe is, because of computational limitations, not accessible to us.[20]

The deeper problem concerning implicit knowledge is that the implicit/explicit knowledge distinction does not track the accessible/inaccessible information distinction. For instance, some explicitly stored information is not accessible in various circumstances. "Repressed" memories might be difficult if not impossible to recover, and some information might be suppressed when various cues are given. Of course, some explicitly stored information *is* accessible to cognition. The distinction between explicit and implicit knowledge does not inform the question of informational access and availability.

3.4.1 Access to Information

Why is accessible information the key to the problem of logical omniscience? Consider two of the cases I previously discussed. First, the student attempting the proof competently reached line n of the proof but couldn't "see" the connection between

[19]Stalnaker (1999), p. 254.
[20]Stalnaker (1999), p. 251.

what he had shown and what he was trying to show.[21] Plausibly, an explanation for his failure to see the connection was a failure of access to information. We can suppose that the student had been taught the particular rule for how to finish the proof, and could typically do so. For whatever reason, however, on this occasion the student could not recognize that "the pieces fit together". Such a failure of recognition is plausibly understood as a failure of access. The second case concerns the familiar story of chess performance. Typical subjects can only see several moves ahead in chess. The standard story is that working memory and computational power are finite and, beyond some threshold, there is no information to access.

3.4.2 Stalnaker on Knowledge

I agree with Stalnaker's general assessment of the problem of logical omniscience, but I am unsure about several of his particular proposals. First, Stalnaker has an external conception of knowledge, one essentially connected with action. As he says,

> Very roughly, I know whether p if I have the capacity to make my actions depend on whether p.[22]

First, this view has some initial plausibility. I know that there is a coffee cup to my right, and I can integrate this information into my planning. That is, I can *successfully* implement the plan "pick up coffee cup when it is to my right" such that I pick up the cup when it is to my right and I do not attempt to pick up the cup when it is absent. In a way, my successful interaction with the cup shows that I am appropriately discriminating or sensitive to the information "there is a cup to my right".[23]

---

[21] See page 71.

[22] Stalnaker (1999), p. 254.

[23] Suppose, instead, that I announce that I have the plan to pick up the coffee cup when it is to my right, and that the coffee cup *is* sitting to my right, but I do not move to pick it up. In this case it seems reasonable to judge that I *do not* know that the coffee cup is to my right.

But Stalnaker also acknowledges deficits in this proposal. He considers a case of a chess player who is able to access information for the purpose of selecting a move but cannot answer questions about the selection. It is plausible that there is a sense in which she knows p (she can select right moves) but also a sense in which she doesn't know p (she can't explain why she did what she did). He then remarks that "the accessibility of knowledge and belief can be understood only relative to the actions they are being used to guide".[24]

It is not perfectly fair to criticize Stalnaker's account of knowledge that he proposes in "The Problem of Logical Omniscience, I" and "The Problem of Logical Omniscience, II" because he doesn't offer a fully worked out theory. But, nevertheless, his rough proposal features prominently in his motivating examples and subsequent discussion. I want to offer several worries for his account before discussing one of the lessons he draws from the problem of logical omniscience.

It seems that I have the capacity to make my actions depend on whether I am unconscious. For instance, I perform the action "prove the Pythagorean theorem" only if I am conscious (presumably I can't do this while unconscious). But I don't know that I am unconscious when I am. As well, most of my actions are dependent on whether my blood glucose level meets and does not exceed some interval (I can't perform an action otherwise). And most of my actions are dependent on some particular sequence of neural firings (or activation in various brain regions).[25] But I don't know whether my blood glucose is in some region and I don't know whether the relevant sequence of neural events has occurred. Like many forms of externalism about knowledge, this account also seems to attribute too much knowledge to subjects.

Second, I am unsure of one particular lesson Stalnaker draws from the problem of logical omniscience. He says,

[24]Stalnaker (1999), p. 254.

[25]Similar examples are easy to generate.

> The reason we idealize in our logics of knowledge and belief is because we have a much clearer conception of implicit knowledge and belief—the information or informational content that we store—than we do of accessible knowledge and belief—the information and belief that is available to guide behavior.[26]

First, Stalnaker's explanation seems to get the order reversed. The historical development of epistemic logics didn't begin by trying to capture a notion of implicit knowledge that had strong closure properties. Hintikka's (1962) insight was to see a commonality between knowledge and necessity: to know p is to exclude all non-p possibilities. Hintikka is often viewed as following G.H. Von Wright's (1951) suggestion that epistemic logic is a branch of modal logic. This characterization, on its own, leads to omniscience properties. For instance, if S knows that p (and, hence, excludes all non-p possibilities) and p entails q, then all non-q possibilities have already been excluded because, in this case, every non-q possibility is a non-p possibility. Second, as I discussed in Chapter 1, all logics of knowledge must idealize to some degree, even ones that attempt to capture non-implicit knowledge. Logics relate structural properties or conceptual relationships, but are able to do so because they isolate critical features of the subject matter and ignore or leave out irrelevant features. For instance, epistemic logics do not typically include a temporal parameter though all agents must reason in time. Many epistemic logics do not typically include a representation of the dynamics of inference (though dynamic logics do). So, Stalnaker cannot say that "the reason we idealize in our logics... is because we have a much clearer conception of implicit knowledge...", because there is no choice whether we idealize when we attempt to give a logic—all logics are idealizations.

Beyond these remarks, I wish to dispute that we do not have a workable conception of accessible knowledge, one that can aid the development of "realistic" epistemic logics. Stalnaker seems to take the pessimistic conclusion that all epis-

[26]Stalnaker (1999), p. 254.

temic logics will be of ideal knowers.[27] But there are models of knowledge that attempt to strip away some of the unreasonable and ideal assumptions present in typical epistemic logics. The best of these models, one I will propose in the next section, looks to cognitive science to help inform a conception of accessible knowledge. We cannot discover these features from the armchair, so this framework will draw on empirical results. This is fitting because the very observation that motivated the problem of logical omniscience began with the empirical discovery of our computational limitations.

3.5 Two Models of Access and Awareness

Recall the fictional example of Sarah suffering from "Sarah's syndrome" I mentioned in section 3.3.1. This syndrome was syntactic in nature and prevented her from being able to process the left disjunct as an independent formula or sentence when she considers a disjunction. Suppose that Sarah's inability to process the left disjunct renders the content of that disjunct inaccessible to cognition. What effect might such a cognitive deficit have on Sarah's deductive ability?

Consider a toy model. Suppose Sarah has two distinct sentences in her vocabulary (individuated syntactically): $\neg\psi, \varphi \vee \psi$. Suppose that Sarah knows all sentences in her vocabulary. That is, Sarah knows p if and only if p is in her vocabulary. Further, suppose that Sarah is a paragon of rationality and attempts to obey the following rule for adjusting her vocabulary (disjunctive syllogism):

> (DS) If $\neg\sigma$ and $\tau \vee \sigma$ are in your vocabulary, then add τ to your vocabulary.

[27] See Stalnaker (1999), p. 245: "perhaps the best we can do is to get a logic of knowledge of an idealized knower, or of knowledge in some special idealized sense". Here he is referring to knowers with unbounded computational abilities and memories, not that all logics must idealize to some degree.

Ever the diligent logician, Sarah attempts to conform to (DS).[28] She finds $\neg\psi$ in her vocabulary and then finds and processes $\varphi \vee \psi$. Yet, due to her syndrome, processing $\varphi \vee \psi$ renders her unable to process and add φ. To make the model more realistic, suppose Sarah becomes "blind" to instances of φ on its own (when she processes $\varphi \vee \psi$). The rule DS then appears to Sarah as

> (DS′) If $\neg\sigma$ and $\tau \vee \sigma$ are in your vocabulary, then add to your vocabulary,

an incomplete version of (DS).

Given Sarah's affliction we may ask whether her knowledge is closed under disjunctive syllogism, (DS), as we interpret it. The answer is clearly "no". Her syndrome prevents her from processing necessary "ingredients" to satisfy (DS), though it may appear to her, from "the inside", that she is not in violation of any rule.

I will next consider a model of knowledge that is able to address some of Stalnaker's worries about "accessible information" that shares several features with the above syntactic toy model. After I introduce the model I will show how it helps illuminate the connection between available information and closure.

3.5.1 Awareness

I want to embrace Stalnaker's assessment of the problem of logical omniscience but not his pessimism. Stalnaker suggests that the problem of logical omniscience essentially concerns the accessibility of information. He also suggests (pessimistically) that there is little to be said, in principle, of our abilities to access information. I disagree on this last point. While I agree that our abilities to access information is a highly contingent matter, perhaps one best addressed by cognitive scientists,

[28] For consistency, we may suppose that Sarah's bout with her syntactic processing syndrome occurs after she learns of DS.

there are models that can help show the structural relationship between kinds of information access and their corresponding closure properties.

For example, one way in which an agent might not have access to some piece of information σ is when the agent is unaware of σ. Suppose Aubry is unaware of Io, a Galilean moon of Jupiter. Aubry has never heard of Io or thought about any of the moons of Jupiter. Suppose some proposition p follows from what Aubry knows, and that the propositional content of p references Io. For concreteness, suppose that p expresses the proposition "Io is the fifth closet satellite to Jupiter". It seems intuitive to say that Aubry doesn't know p because she is, in some sense, unaware of p. And it also seems intuitive to say that Aubry does not have access to the information that p because she is unaware of p.

These remarks suggest a way of strengthening the formal approach of modeling knowledge. I previously gave the standard formalization of knowledge, and interpreted it in the common way as "knowledge is truth in all accessible worlds". The formalism expresses this property as:

$$(M, s) \models K_i\varphi \text{ if and only if } (M, t) \models \varphi \text{ for all } t \text{ such that } (s, t) \in \mathcal{K}_i.$$

But we have reason to amend this formalism by including an awareness condition. That is, it seems plausible that knowledge requires awareness (of a relevant sort). If Aubry is unaware of p then Aubry doesn't know p.

Consider the following model adapted from Fagin et al. (1995). Begin with an "awareness structure" $M = (S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{A}_1, \ldots, \mathcal{A}_n)$, where the tuple $(S, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n)$ is a Kripke structure for n agents and $\mathcal{A}_i$ is a function that assigns a set of formulas for each agent i with each state.[29]

The function $\mathcal{A}_i$ is the new addition to the standard Kripke model of knowledge. I will have more to say about its interpretation, but, for now, its interpretation can be viewed as somewhat open. View the function $\mathcal{A}_i(s)$ as assigning the set of

[29]My presentation of this awareness model follows Fagin et al. (1995).

formulas that i is aware of at state s. If Aubry is not aware of the planet Io and p expresses the proposition "Io is the smallest of Jupiter's moons", then it may be proper to render the model such that $p \notin \mathcal{A}_i(s)$ at state s. But we could also view $\mathcal{A}_i(s)$ as capturing a notion of access to information. We would then view $\mathcal{A}_i(s)$ as returning the set of formulas that i has access to at s.

These consideration lend naturally to an amended model of knowledge. To update the standard Kripke structure for knowledge, add two new clauses for formulas of the form $A_i\varphi$ ("i is aware of φ") and $X_i\varphi$ ("i knows φ"):

(AW) $(M, s) \models A_i\varphi$ if and only if $\varphi \in \mathcal{A}_i(s)$,

(KN) $(M, s) \models X_i\varphi$ if and only if $(M, s) \models A_i\varphi$ and $(M, t) \models K_i\varphi$.

Condition (AW) captures the idea that an agent i is aware of a proposition p if and only if p is among the set of propositions she is aware of at s. The model specifies this set syntactically. Note that we could also understand (AW) as capturing a notion of access to information.

(KN) provides the amended definition of knowledge. The knowledge operator X is just the familiar operator K with the added condition that i is also aware of φ (or, alternatively, φ is accessible to i). Again, the motivation for (KN) is that, in many cases, it is plausible to say the following:

(1) S doesn't know p because S is unaware of p.[30]

For a concrete instance, I can say of my seven-year-old nephew that he doesn't know that aesthetic properties supervene on physical properties because he has never heard of supervenience relations or aesthetic properties and is, hence, unaware. I can also say of my seven-year-old nephew that he doesn't know that Don Bradman

[30]It also seems correct to say, in many cases, S doesn't know p because S doesn't have access to the information that p, but this is a theoretical consideration, not one based on philosophical intuition.

was Australian because he's never heard of Don Bradman and is unaware of Don Bradman.

More generally, why think (1) is plausible? One reason references epistemic possibility. Michael Huemer (2007) suggests that attributions of epistemic possibility presuppose a kind of awareness.[31] Simply, S's not being aware of p makes it that p is not epistemically possible for S. But if p is not epistemically possible for S then S can't know p.

Before I discuss different interpretations of $\mathcal{A}_i$, I want to show how various closure principles do not hold for the knowledge operator X. Recall the closure principles (CLI) and (CLE) from page 68. (CLI) and (CLE) need not hold for models of knowledge in awareness structures. Note that in the following discussion I represent "S knows that p" as "$X_s p$".

First, consider (CLI) (closure under logical implication). Suppose S knows φ and that φ logically implies ψ. Must it be the case that S knows ψ? No. As it was assumed, S knows φ, so $\varphi \in \mathcal{A}_s(s)$ by (KN). But it need not be the case that $\psi \in \mathcal{A}_s(s)$, so it need not be the case that S knows ψ (that is, it is consistent with our assumptions that $\psi \notin \mathcal{A}_s(s)$).

Next, consider (CLE) (closure under logical equivalence). Suppose that S knows φ and that φ and ψ are logically equivalent. Must it be the case that S knows ψ? Again, no. From the assumption, S knows φ, so $\varphi \in \mathcal{A}_s(s)$ by (KN). But it need not be the case that $\psi \in \mathcal{A}_s(s)$, so it need not be the case that S knows ψ.

Beyond (CLI) and (CLE), consider a stronger version of (CLI):

> (CKLI) If S knows φ, and if S knows that φ logically implies ψ, then S knows ψ.

This principle, closure under known logical implication, is surely more plausible than (CLI). In fact, (CKLI) is close to principle (B) from section 3.2. Yet (CKLI)

[31] See Huemer (2007), p. 122–123.

is not true in all awareness structures. Such a result is easy to see from the above discussion of (CLI) and (CLE). Suppose that S knows φ and also that S knows that φ logically implies ψ. It need not be the case that S knows ψ because it need not be the case that $\psi \in \mathcal{A}_s(s)$, even though '$\varphi \Rightarrow \psi$' $\in \mathcal{A}_s(s)$. This is possible, in an awareness structure, because, syntactically, '$\varphi \Rightarrow \psi$' and 'ψ' are different formulas.

I think it is a virtue of this model of knowledge that (CKLI) is not always true. As I gave the example earlier, there are cases when knowledge is not closed under *known* entailment (or logical implication). Again, suppose that Sarah comes to know axiom p from testimony by her college number theory professor. Later in the semester Sarah notices that p entails q (by a complicated proof), but she doesn't notice that her proof is related to what she learned before from her professor and does not come to know q. One explanation of this case is that Sarah's failing is of not noticing the "connection" between her proof and what she may properly come to know (that is, q). Interpreting $\mathcal{A}_i(s)$ as of "access", it is not unreasonable to think that Sarah did not have access to q.

Properties of Awareness Structures

In the above example, Sarah acted imperfectly. A better agent, perhaps, would not make the kind of mistake of which Sarah is guilty. Awareness structures provide enough flexibility to model these kinds of mistakes.

From the case, Sarah was aware of '$\varphi \Rightarrow \psi$' but she was not aware of 'ψ'. When $\mathcal{A}_i$ is interpreted as that of access such a phenomenon may seem commonplace. I may remember that taking my keys entails that Jo cannot get into the office and remember that I took my keys, but fail to see that Jo cannot get into the office because I don't consider Jo. But such a phenomenon may seem less realistic when $\mathcal{A}_i$ is interpreted as that of awareness. Is it possible to be aware of '$\varphi \Rightarrow \psi$' but not be aware of 'ψ'?

Awareness structures allow tremendous modeling flexibility. One understanding

of awareness shares features with a notion of recognition. One may recognize a painting but not recognize a small sub-region of the painting when only the sub-region is presented. One may recognize a person's face, but not recognize their eyes alone (for instance). In this regard, it is not absurd to think that an agent might be aware of a formula but not its parts. Nonetheless, if it seems unreasonable that an agent could be aware of '$\varphi \Rightarrow \psi$' but not its constituent parts 'φ' and 'ψ', we could impose the following restriction on $\mathcal{A}_i$:

> *Entailment Subformula Closure*: if $\varphi \Rightarrow \psi \in \mathcal{A}_i(s)$ and ξ is a subformula of $\varphi \Rightarrow \psi$, then $\xi \in \mathcal{A}_i(s)$.

If $\mathcal{A}_i$ were restricted by Entailment Subformula Closure then (CKLI) holds. To see why, suppose that S knows φ and S knows that $\varphi \Rightarrow \psi$ at state s and that $\mathcal{A}_i$ is restricted by Entailment Subformula Closure. Because of the entailment, ψ is true at s, and $\psi \in \mathcal{A}_i(s)$ by Entailment Subformula Closure. So S knows ψ.[32]

It should be clear that there is flexibility in the properties we give to $\mathcal{A}_i$. How we understand $\mathcal{A}_i$ depends on our purposes. Perhaps, for relatively ideal agents, we should view their awareness as satisfying Entailment Subformula Closure. But there are many other such restrictions. Consider the following:

> (A1) General closure under subformulas: if $\varphi \in \mathcal{A}_i(s)$ and ψ is a subformula of φ, then $\psi \in \mathcal{A}_i(s)$.
>
> (A2) Generation by primitives at a state s: for each s, S is aware of a set Ψ_s of primitive propositions at s, and $\mathcal{A}_s(s)$ is generated by the closure of Ψ over the connectives "$\wedge$", "$\neg$", and operators "K", "X", and "A".
>
> (A3) Generation by primitives: S is aware of the same set Ψ of primitive propositions at every state s, and $\mathcal{A}_s(s)$ is generated by the closure of Ψ over the connectives "$\wedge$", "$\neg$", and operators "K", "X", and "A".[33]

[32] Note that Entailment Subformula Closure *does not* satisfy (CLI) in all awareness structures.

[33] Note that (A3) necessarily satisfies the condition that if $(s,t) \in \mathcal{K}$ then $\mathcal{A}_s(s) = \mathcal{A}_s(t)$.

(A4) Generation by primitives of length n: S is aware of the same set Ψ of primitive propositions at every state s, and $\mathcal{A}_s(s)$ is generated by the closure of Ψ over the connectives "$\wedge$", "$\neg$", and operators "K", "X", and "A", where the "length" of each element of $\mathcal{A}(s)$ does not exceed k primitive propositions and l connectives and m operators, with $k + l + m = n$.

(A5) Awareness of awareness: if $\varphi \in \mathcal{A}_i(s)$ then $A\varphi \in \mathcal{A}_i(s)$.[34]

(A6) Symmetry "$\vee$": For all $s \in S$, $(\varphi \vee \psi) \in \mathcal{A}_i(s) \Leftrightarrow (\psi \vee \varphi) \in \mathcal{A}_i(s)$.

(A7) Symmetry "$\wedge$": For all $s \in S$, $(\varphi \wedge \psi) \in \mathcal{A}_i(s) \Leftrightarrow (\psi \wedge \varphi) \in \mathcal{A}_i(s)$.

(A8) Negation complete: if $\varphi \in \mathcal{A}_i(s)$ then $\neg\varphi \in \mathcal{A}_i(s)$.

(A9) Knowledge of awareness: an agent knows of which formulas he is aware when it is the case that if $(s, t) \in \mathcal{K}$, then $\mathcal{A}_i(s) = \mathcal{A}_i(t)$.

(A10) Full awareness: for any $\varphi \in \Phi^{K,A,X}$, $\varphi \in \mathcal{A}_i(s)$.

(A11) No awareness: for any $\varphi \in \Phi^{K,A,X}$, $\varphi \notin \mathcal{A}_i(s)$.

(A12) No awareness of unawareness: if $\neg A\varphi \in \mathcal{A}_i(s)$ then $A\neg A\varphi \notin \mathcal{A}_i(s)$.[35]

The knowledge operator X will have different properties depending on the properties (A1)–(A12) given to $\mathcal{A}$. Below I'll discuss how these properties influence X.

At either extreme, both (A10) and (A11) have straightforward implications for X. Suppose (A10) holds for $\mathcal{A}$. In this case, X is then equivalent to K. Suppose $K\varphi$. Trivially, then, $A\varphi$, so it then must be the case that $X\varphi$.[36] And this should

[34]This corresponds to the axiom $A\varphi \Rightarrow AA\varphi$. Such an axiom might be true for agents who, when they are aware of some formula, are aware that they are aware of the formula.

[35]The restriction should actually be more general. The idea is that if S is unaware of some formula, then there can be no positive iteration of awareness of this fact.

[36]The other direction follows from the definition of X.

be unsurprising. The function $\mathcal{A}$ works as a restriction on K (for X), but when $\mathcal{A}$ is unbounded, X is just K.

Next, suppose (A11) holds for $\mathcal{A}$. In this case, the extension of X is empty—at every state s, the set of propositions known by S is empty. That is, when S is not aware of any proposition, S has no explicit knowledge.

The restriction (A9) has more interesting implications. Suppose (A9) holds for $\mathcal{A}$. Then in all awareness models M, the following hold: "$A\varphi \Rightarrow KA\varphi$" and "$\neg A\varphi \Rightarrow K\neg A\varphi$". The first represents knowledge of awareness. Suppose that $A\varphi$ at s, and that (A9) holds. Then in all states t such that $(s,t) \in \mathcal{K}$, it is true that $A\varphi$. So, by definition of K, $KA\varphi$ at s.[37] The second represents knowledge of unawareness. Suppose that $\neg A\varphi$ at s, and that (A9) holds. Then in all states t such that $(s,t) \in \mathcal{K}$, $\neg A\varphi$. So, by the definition of K, $K\neg A\varphi$. While knowledge of unawareness may sound peculiar, note that this only concerns the K operator, the "implicit knowledge" operator. However, $\neg A \Rightarrow X\neg A\varphi$ is satisfiable. One restriction on $\mathcal{A}$ to avoid such results would be to add (A12). For a contradiction, suppose $\neg A\varphi \Rightarrow X\neg A\varphi$ and (A12). Suppose $\neg A\varphi$. Because $X\neg A\varphi$, it must then be the case that $A\neg A\varphi$ (by definition of X). But this conflicts with (A12).

The restriction (A8) that $\mathcal{A}$ be negation complete has the implication that subjects explicitly know all propositional tautologies of the form $(\varphi \vee \neg\varphi)$, when they're aware of one of the disjuncts.[38] Certainly, it is necessarily true that $K(\varphi \vee \neg\varphi)$, because $(\varphi \vee \neg\varphi)$ is true in any world S considers.

The symmetry requirements in (A6) and (A7) are minimal. That is, if (A6) and (A7) are the only restrictions on $\mathcal{A}$ then non-relativized closure property $X\varphi \wedge X(\varphi \Rightarrow \psi) \Rightarrow X\psi$ need not hold. Nor need the non-relativized generalization property hold for X: for all structures M, if $M \models \varphi$, then $M \models X\varphi$.

[37] When A has the property that $A\varphi \Rightarrow AA\varphi$, then it is also true that $A\varphi \Rightarrow XA\varphi$.

[38] Because syntactic equivalence is not semantic equivalence, we also need to assume that if $\neg\neg\varphi \in \mathcal{A}(s)$ then $\varphi \in \mathcal{A}(s)$.

Next, it is important to observe that restriction (A3) entails (A1). If S's awareness is generated by some set Φ of primitive propositions, then necessarily awareness will be closed under subformulas. Suppose it is not closed. Then there would be some subformula ψ of φ such that $\psi \notin \mathcal{A}(s)$ for some state s. But because ψ is a formula, it is composed of n primitive propositions and joined by connectives and operators. Hence, $\psi \in \mathcal{A}(s)$, a contradiction.

Consider restriction (A1). The restriction of closure under subformulas can be axiomatized by the following axioms:

$$A(\neg\varphi) \Rightarrow A\varphi,$$
$$A(\varphi \wedge \psi) \Rightarrow (A\varphi \wedge A\psi),$$
$$A(X\varphi) \Rightarrow A\varphi,$$
$$A(K\varphi) \Rightarrow A\varphi,$$
$$A(A\varphi) \Rightarrow A\varphi.$$[39]

If (A1) holds for $\mathcal{A}$, then a non-relativized version of closure under implication is valid for X: namely, $X\varphi \wedge X(\varphi \Rightarrow \psi) \Rightarrow X\psi$ is valid. In this case, we need not add the further clause $A\varphi$ because (A1) provides that $A\varphi$ from the fact that $X(\varphi \Rightarrow \psi)$. However, even when (A1) holds it is not the case that subjects S explicitly know all valid formulas.

Interpretation of $\mathcal{A}_i$

The function $\mathcal{A}_i$ is sentential in its operation. It associates a set of formulas at each state $s \in S$. Above, I informally interpreted the set $\mathcal{A}_i(s)$ as the formulas the agent is aware of at state s. I also mentioned that there is some evidence between our practice of knowledge attribution and awareness attribution. If we know that

[39]Changing the "$\Rightarrow$" to "$\Leftrightarrow$" would provide for the restriction (A3), awareness generated by a set of primitive propositions.

Sarah is unaware of Io, that she has never heard of the moon Io, then we shouldn't attribute knowledge of the proposition "Io is the fifth closest satellite to Jupiter" to Sarah. If questioned, we could properly respond that Sarah can't know that Io is the fifth closet satellite to Jupiter because she's never heard of the moon.

However, the expression 'aware of' is somewhat flexible in its meaning, and we can use its flexibility to model slightly different cognitive phenomena. A weak understanding of awareness might suggest something close to "ever heard of". So, for example, an agent is weakly aware of a formula ψ if there was ever a time the agent ever considered, thought of, heard, processed, or reflected on ψ. Very informally, agent S is weakly aware of ψ if there was ever a time t when ψ was "on S's radar". On this rendering of "aware of", once S is weakly aware of ψ, S is always weakly aware of ψ.

Yet, there are other interpretations of "aware of", which may be appropriate for different purposes. A more strict interpretation of "aware of" might be close in meaning to "can process" or "available". So, for example, an agent S might be strictly aware of ψ when S is weakly aware of ψ *and* S can recall ψ or reason with ψ or otherwise cognitively process ψ. Again, very informally, an agent S might be thought of as strictly aware of ψ when S can cognize with ψ. Never having heard of ψ is one bar against being strictly aware of ψ (in virtue of not being weakly aware of ψ), but not being able to remember ψ is another bar against being strictly aware of ψ (because such information is not "available").

For my purposes, I only wish to note the theoretical flexibility that awareness structures can provide to model various cognitive phenomena. In essence, they provide a theoretical degree of freedom. A natural yet broad interpretation of the set $\mathcal{A}_i(s)$, when coupled with the operator X, is of a kind of "cognitive bound". Think of $\mathcal{A}_i(s)$ as an agent's "cognitive vocabulary" at s, the mental items at her disposal at s. Because agents can be cognitively bounded in various ways, the set $\mathcal{A}_i(s)$ can plausibly be employed to model these various cognitive limitations.

Various restrictions or properties of $\mathcal{A}_i$ (as I mentioned above in (A1)–(A12)) will be appropriate depending on the intended model.

Below, in section 3.7.1, I discuss the sense in which what we know is in many ways indefinite. In the subsequent discussion I will make use of the flexibility of "aware of" in the ways I've just described.

3.5.2 Local Reasoning and Access

In standard epistemic logics (with semantics given by Kripke structures), each agent has one accessibility relation, which describes the states that are epistemically possible from some state s. The intended interpretation of the accessibility relation, $\mathcal{K}_i$, is of epistemic possibility. That is, $\mathcal{K}_i(s) = \{t : (s,t) \in \mathcal{K}_i\}$ picks out the states t that i cannot differentiate between, given her evidence, knowledge, or justified beliefs. For example, given what I know right now, I cannot tell the difference (tell which is the actual state) between the state of the world being such that it is raining in Paris right now and the state of the world being such that it is not raining in Paris right now. Given what I know, it is possible that it is raining in Paris. To contrast, given what I know right now I *can* tell the difference between the world in which my cat is sleeping on the floor nearby and one where my cat is not sleeping on the floor. Because I know my cat is sleeping on the floor nearby, the actual state must include him sleeping on the floor. It is not possible that my cat is not sleeping on the floor.

The above case is ideal in at least this respect: given S's accessibility relation, the epistemically possible states for S remain fixed. But sometimes we make mistakes about epistemic possibility that cannot be captured by Kripke structures. Consider a familiar example. Suppose you know that you play bridge on Mondays at Martha's house. Also, suppose that you learn that next Monday Martha will be out of town. Ideally, you should put these two pieces of knowledge together and infer that you will not be playing bridge next Monday. But sometimes we make errors of the following

sort: somehow we're able to entertain these two separate thoughts (perhaps not at the same time), "I play bridge on Monday" and "Martha will be out of town", yet never reconcile them. Informally, it is almost as if we have two different "frames of mind". For part of the day we think that we have bridge, for the other part we take notice that Martha is out of town.

Consider another example. I can imagine a well meaning politician, overly eager to please, who, on some occasions professes that we ought to balance the budget and on other occasions professes that we need to increase spending on education. Let's suppose that the politician is not a liar, but is making a more subtle mistake. Suppose that she tries to reason well, and when she considers whether to balance the budget she only considers possible states s_1 and s_2, which favor balancing the budget. Yet, when she considers whether to increase education spending she only considers possible states s_3 and s_4, which favor an increase in spending. She acts as if she has two frames of mind, complete with different possibilities for each.

To capture such a phenomenon, extend the standard model of knowledge with multiple accessibility relations, one for each "frame of mind" for the agent. Call $M = (S, \pi, \mathcal{C}_1, \ldots, \mathcal{C}_n)$ a "local reasoning structure", where S is a set of states and $\mathcal{C}_i(s)$ is a non-empty set of subsets of S for each agent i.[40] The idea is that each $\mathcal{C}_1, \ldots, \mathcal{C}_n$, where $\mathcal{C}_i(s) = \{T_1, \ldots, T_n\}$, represent different "frames of mind", concomitant with a stock of possibilities—sometimes i takes the set of possibilities to be T_1 at s, and sometimes i takes the set of possibilities to be T_2 at s (and so forth).[41] Perhaps, when the politician appears before a business association she only considers as possible ways the world might be that support balancing the budget. But when the politician appears before young college students she only considers as possible

[40]See Fagin et al. (1995) for their presentation of local reasoning structures. Fagin and Halpern (1988) relate that their "frame of mind" approach was influenced by Stalnaker (1987).

[41]Note that $M = (S, \pi, \mathcal{C}_1, \ldots, \mathcal{C}_n)$ represents a multi-agent local reasoning structure. For a single agent let $M = (S, \pi, \mathcal{C})$, where $\mathcal{C}(s) = \{T_1, \ldots, T_n\}$ represents n "frames of mind" for the agent.

ways the world might be that support spending on education. Clearly, such behavior is not ideal—we would hope that the politician would notice her inconsistency and resolve the conflict. Yet, such phenomena occur and say something about our powers of reasoning and have implications for human knowledge.

How might we model knowledge in local reasoning structures? Modify the definition for knowledge in Kripke structures in the following way (with Z acting as a new knowledge operator):

> $(M, s) \models Z_i\varphi$ if and only if there is some $T \in \mathcal{C}_i(s)$ such that $(M, t) \models \varphi$ for all $t \in T$.

So far, the interpretation of Z is *not* of a standard notion of knowledge. Without any restrictions on $\mathcal{C}_i(s)$ formulas such as "$K_ip \wedge K_i\neg p$" are satisfiable in local reasoning structures, because the agent may know p in one frame of mind and know $\neg p$ in another frame of mind. That is, without further restrictions the operator Z is not guaranteed to be factive. To remedy such a defect, we may assume that the actual state s is a member of every $T \in \mathcal{C}_i(s)$. The result of this assumption is that $Z\varphi \Rightarrow \varphi$ is correspondingly valid in every local reasoning structure.

As should be clear, various closure properties do not hold for local reasoning structures. Consider the following closure principle (closure under material implication):

> (CMI) If S knows φ and S knows $\varphi \rightarrow \psi$, then S knows ψ.

To see why (CMI) fails, suppose that S knows φ and that S knows $\varphi \rightarrow \psi$. This may be true if S knows φ in frame of mind T_1 and knows $\varphi \rightarrow \psi$ in frame of mind T_2, but there is no frame of mind where S knows both these propositions. In this case, S does not know ψ, so CMI fails.

Next, it is straightforward to show that knowledge need not be closed under conjunction (CC) in all local reasoning structures:

(CC) If S knows φ and S knows ψ, then S knows $\varphi \wedge \psi$.

Again, S may know φ in frame of mind T_1 and S may know ψ in frame of mind T_2, but there need not be some frame of mind T_3 where S knows both these propositions, so, in this case, S does not know $\varphi \wedge \psi$.

Though local reasoning structures have shortcomings as a model for knowledge (I will address some concerns below), they provide some insight into a workable notion of information access with a logical pedigree, one Stalnaker thought was not achievable. The various members $T \in \mathcal{C}_i(s)$, interpreted as "frames of mind", may be thought to specify a kind of cognitive access. Consider, again, the phenomenon of considering some possibilities in one frame of mind, and other distinct possibilities in another frame of mind. Terry Horgan gave me an interesting example of this. Paraphrasing his example, he suggested that,

> Many philosophers of language have concluded that belief contexts of utterance are intentional contexts. Yet, many technically minded epistemologists have sought to understand rational degree of belief or "rational credence" as a kind of probability. However, few have noticed the need to reconcile these views (when, no doubt, these technical philosophers are conversant with philosophy of language).

I want to suggest that local reasoning structures can help model this situation and clarify its consequences. It is plausible to assume that in this case, considering technical issues lends one to a particular frame of mind, and considering issues of language lends one to a different frame of mind. That is, when one is engaging with technical issues such as subjective probability it seems plausible that various kinds of considerations appear as most relevant or salient, such as the existence of various representation theorems, for instance. Yet when one is engaging with issues of language different considerations appear as most relevant.

With the example in mind, suppose that philosopher S has two frames of mind, T_1 (the probability mode) and T_2 (the philosophy of language mode). Suppose that the set of possibilities from T_1 are such that S knows φ, and the set of possibilities from T_2 are such that S knows $\varphi \rightarrow \psi$, with the intended interpretation of φ as "I am investigating a belief context" and ψ as "I am investigating an intentional context". Suppose, further, that S is in the probability mode of thinking. What is intuitively going wrong in this case is that S does not "see" the connection between her beliefs. This presents a case where closure under material implication should not hold. Local reasoning structures show how this is possible, for S may know φ, may know $\varphi \rightarrow \psi$, but not know ψ.

It is natural to make further assumptions about such a case. We may suppose that when S is in one frame of mind T_1, other frames of mind T_n are not available to S (that is, the set of possibilities in T_n are not available). In this way, local reasoning structures may provide a notion of information access. State s is not possible for S when $s \in T_2$ but S is in frame of mind T_1 and $s \notin T_1$.

Yet, local reasoning structures have limitations. Though knowledge is not closed under material implication and conjunction, knowledge is closed under valid formulas and logical implication. To see why (CLI) holds, suppose that S knows φ and that $\varphi \Rightarrow \psi$. If S knows φ then there must be some frame of mind T_n such that every accessible state t has it that φ. But because $\varphi \Rightarrow \psi$, every accessible state t also has it that ψ. So S knows that ψ.

There are two responses to this result. First, one may say that local reasoning structures are still idealizations (every epistemic logic will have idealizations), yet they are improvements in the direction of realistic agents. But, second, one may also merge local reasoning structures with awareness structures to avoid principles such as (CLI) and (CKLI). What is potentially illuminating about local reasoning structures will be preserved when merged with suitable awareness structures.

3.6 Closure, Revisited

I previously gave counterexamples to closure principles (A) and (B) (see pages 65 and 66). Closure principle (C) is a bit more difficult because what is required for "competent deduction" is complicated. I'll say more about competent deduction below, but, for now, it is plausible that the example of the student failing to see the connection between his proof and what he is trying to prove provides a counterexample to (C).

Hawthorne (2005) also acknowledges deficiencies in (C) and proposes a fix:

> (D) If one knows p and competently deduces q from p, thereby coming to believe q, while retaining one's knowledge that p, one comes to know that q.[42]

Closure principle (D) addresses issues of memory ("...while retaining one's knowledge..."), similar enough, perhaps, to a case of a student who forgets an axiom from her number theory professor. But (D) is still not satisfactory. An agent may satisfy all the conditions in the antecedent in (D) but not "see" the connection between her proof, her original knowledge, and what is proved. Consider, as a partial remedy, the following:

> (E) If one knows p and competently deduces q from p, thereby coming to believe q, while retaining one's knowledge that p, and learns of no undefeated defeater for q in the process, and one "sees" the connection between one's deduction, one's original knowledge that p, and what is proved, one comes to know that q.

This modified closure principle has advantages over principle (D). Kvanvig (2006) proposed including the clause regarding "undefeated defeaters for q" and it is a

[42] See Hawthorne (2005), p. 29.

sensible addition. As he remarks, "rarely do we learn nothing in the process of deducing a claim other than the claim itself". As well, suppose the additional clause requiring that the agent "sees" the connection between the proof and original knowledge handles the cases offered against principle (C).

Before addressing the adequacy of (E), I want to point out an important feature of the kinds of response given to proposed counterexamples to these closure principles. The proposed counterexamples aim to show that the conditionals representing the closure principles are false, and the subsequent modifications and emendations to the principles all work by strengthening the antecedent in the conditional. That is, the modifications which lead to (B), (C), (E), and (D) all effectively avoid their particular counterexample by endowing the fictional agent with a stronger epistemic position (in a manner of speaking). For instance, closure principle (A) is false because we aren't aware of all the deductive consequences of our knowledge. Closure principle (B) is an improvement over (A) for two reasons. First, principle (B) limits the range of applicability to known entailment (rather than entailment). Second, principle (B) adds the assumption that the agent *knows* the entailment holds. Principle (C) further adds the assumption that the agent competently deduces the proposition (rather than only knowing that the entailment holds). Principle (D) adds the assumption that the agent doesn't forget her original knowledge. And so on.

One way of understanding this exchange of closure principle, counterexample, and modification to closure principle, is as of sequential strengthening of the agent's cognitive powers and epistemic position. Each modification to the closure principle makes the antecedent more cognitively demanding and thereby more difficult to satisfy. Such modifications recall Stalnaker's discussion of the problem of logical omniscience. For, a plausibly true closure principle is the following:

> (F) If one knows p and knows that p entails q and is perfectly logically omniscient and has no cognitive limitations, then one knows q.

Closure principle F seems true.[43] One might argue whether (F) is trivially true (at least, (F) is not a tautology nor an a priori truth), but this is to miss the point. The point is that principle (F) is unacceptable because it is uninformative. Humans cannot satisfy the antecedent condition so the principle does not explain the platitude that deduction is a way of extending *our* knowledge.

Stalnaker's (1999) discussion of the problem of logical omniscience makes it clear that, yes, logically omniscient and cognitively unbounded agents know all the deductive consequences of their knowledge. Stalnaker shows that the essential problem of saying something informative about closure must address the issue of access to information, one that he thinks is too complicated and messy to be informed by logic. I am more sanguine about the situation. While I agree that one cannot say something informative about the issue of access to information by logic *alone*, I think there is room for models of knowledge that incorporate empirical facts from the empirical domain. The kinds of models I envision were sketched in section 3.5 above.

I can give the above remarks in another way. The platitude that deduction can extend *our* knowledge is both defended and explained only when the focus is on *us* as human knowers. That is, the platitude essentially concerns realistic (non-ideal) agents. Our cognitive limitations are then key. Surely, lapses in memory destroy knowledge. But missing various connections between deductive relations and proper inference also destroys knowledge. Models of awareness and access to information can help show the interrelation between these cognitive failures and their consequences for closure.

[43]Fred Dretske (2005) would, no doubt, challenge this principle when q is a "heavyweight" proposition.

3.6.1 Competent Deduction and "Seeing" Connections

Several authors have gestured at the correctness of principles like that of (C) and (D) (especially Williamson (2000) and Hawthorne (2005)) so it is worthwhile discussing what is meant by competent deduction. Naturally, "competence" evokes a competence/performance distinction, but this is not quite correct for purposes of closure. Typically, competence/performance distinctions are employed to explain performance failures when an agent "really knows better" (in some sense). For instance, fluent speakers of a language make occasional grammatical mistakes. These mistakes do not impugn the tacit knowledge an agent actually has of grammaticality, only that on a particular occasion such knowledge was not utilized or applied (perhaps for cognitive failures such as memory limitations, etc.).[44] But, crucially, "competent deduction" (as used in closure principles) must be *successful* deduction to impart knowledge. Presumably one could competently deduce a proposition that does not follow—but then one wouldn't know the result of the deduction. In this way, competent deduction is *not* just the familiar competence/performance distinction.

What, then, is competent deduction? I want to suggest that the expression itself is not that helpful. First, 'competent deduction' is not an ordinary expression, so we cannot look to ordinary use to provide constraints (as we do when we note that 'knows' must be factive). Because we know the job it is needed to perform, competent deduction should rule out lucky deduction, as when a student just happens to correctly use several rules of proof. Yet it is less clear whether competent deduction provides that an agent "see" the connection between the deduction, what was deduced, and one's original knowledge. Plausibly, an agent can competently deduce q from p by expertly following the rules of proof, but not "see" how such facts "fit

[44]See Chomsky (1965), p. 3, for one of the first uses of this distinction: "an ideal speaker-listener... is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, ...".

together" (I gave such a case in section 3.3.1). A mathematician might competently produce an exceedingly long proof involving, say, four hundred steps, and not "see" that her proof demonstrates that one proposition follows from another. But, again, there are no constraints on usage that force one to say this.

As I previously suggested, the closure of knowledge fundamentally concerns the availability or access of information. Both "competent deduction" and "seeing connections" of deductive consequences can be viewed as attempting to provide for the availability or access of the relevant information. These expressions on their own, however, are not informative. I introduced two models of awareness and access to information to help show the interrelation between access, awareness and closure. These models provide the first steps to informative principles that can explain the platitude that deduction can extend our knowledge while, at the same time, respect our cognitive limitations.

3.7 Remarks on Skepticism

Intense philosophical interest in closure principles began with Fred Dretske's (1970) paper "Epistemic Operators". In his paper, Dretske famously argued against the closure of knowledge, helped in part with an example of a zebra at the zoo. Dretske showed how denying closure could potentially offer a response to one kind of skeptical argument. For example, consider the following argument:

(S1) If I know I have a hand, then I know I am not a brain in a vat.

(S2) I do not know I am not a brain in a vat.

(S3) So, I don't know I have a hand.

Plausibly, (S1) is true by some closure principle, because having a hand supposedly entails that one is not a brain in a vat.[45] (S2) seems intuitively true—most people

[45] I say more about this supposed entailment below.

seem happy to admit that this isn't the kind of thing we can know. (S3) deductively follows from (S1) and (S2). The skeptic can then take advantage of closure by noting that most everything we think we know seems to entail remote propositions we routinely take ourselves to not know. That there is a coffee cup on my desk entails that I'm not being deceived by an evil genius, and, plausibly, I cannot know *that*. Dretske's argument against closure blocks this particular avenue for the skeptic. If closure is false, we have a general reason to think that (S1) is false, and, hence, do not have reason to accept (S3). If Dretske is correct then I *can* know that I have a hand *and* not know that I am not a brain in a vat, and hold these propositions consistently.

I will address Dretske's argument below. But, for now, I will give a few remarks about skeptical arguments that employ closure principles and show how what I've previously said provides grounds for a different kind of response to skepticism.

3.7.1 The Indefiniteness of Knowledge and Skeptical Entailment

Skeptical arguments that tacitly employ closure principles (such as (S1) above) rely on the existence of entailments between what may be called "ordinary propositions" such as "I have a hand" and what may be called "skeptical propositions" such as "I am not a brain in a vat". It is, however, important to be clear about this relationship.

Strictly speaking, the proposition "I have a hand" does not *logically* entail that "I am not a brain in a vat". As a matter of logic, one *could* be a brain in a vat with hands.[46] The proposition "I have a hand" entails that "I am not handless" (by double negation), and there are many ways in which someone may be handless. One may be a handless banker, one may be a handless computer engineer, or one may be a handless brain in a vat. So, "I have a hand" *does* entail that "I am not a handless brain in a vat", because "I have a hand" entails that "I am not handless"

[46]There is another interpretation of what it means to be a brain in a vat and I will discuss this next.

in any way.

To see this more clearly, let h denote the proposition "I have a hand". By the truth-functional meaning of "$\vee$", the following holds:

(i) $h \Rightarrow (h \vee \neg p)$,

for any proposition p. But $h \vee \neg p$ is logically equivalent to $\neg(\neg h \wedge p)$. So, the following relationship is equivalent to (i):

(ii) $h \Rightarrow \neg(\neg h \wedge p)$,

for any proposition p. From (ii) it is clear why "I have a hand" entails that "I am not handless and a brain in a vat" (being careful to interpret the negation as having scope over the entire conjunction).

What (ii) shows is that an ordinary proposition like h logically entails an infinite number of propositions (which is not surprising). Again, if one has hands then one is not a handless astronaut (one is not handless and an astronaut), and one is not a handless philosopher, and one is not a handless cartographer, and so on. In this way, statement (ii) can be understood as a schema, having a variable or placeholder in the position of the right conjunct such as

(iii) $h \Rightarrow \neg(\neg h \wedge \_\_\_)$,

where '___' represents the inclusion of any proposition.

Statement (iii) shows how this entailment is "open-ended" or unbounded. In light of my remarks about unawareness, might a human agent be able to fully enumerate the open-ended entailment represented by (iii)? No. Because human agents face computational limitations, their view of the world is necessarily bounded or closed—there are always contingencies and propositions we do not consider, and contingencies and propositions we cannot consider. Human agents *are* able to provide an *incomplete* list when filling out (iii): having hands entails that one is not a handless skydiver, or a handless hairdresser, or a handless sea anemone...

Now, one may interpret "I am not a brain in a vat" to mean "I am not deceived". If there is an entailment from "I have a hand" to "I am not deceived" then it is not perfectly straightforward. Surely, the proposition "I have a hand" does not (on its own) entail "I am not mistaken about having hands" because one may have hands and think that one does not (by way of delusion or hallucination), and hence be mistaken.[47] At any rate, (ii) provides a way of understanding "I am not deceived". Let p be the proposition "I am not mistaken about having hands". By (ii), "I have hands" entails that one is not handless and not mistaken about having hands (again, being careful to interpret the negation as having scope over the entire conjunction). Yet, seen in this way, being mistaken about having hands while handless is just another way of being handless.

These remarks, in conjunction with my discussion of models of unawareness for realistic, non-ideal agents, suggest that agents represent the world in incomplete terms. Perhaps, if they view themselves as having hands, they view themselves as not being a handless gorilla, but they may not have considered themselves as not being a handless brain in a vat, because they never considered there being any envatted brains. Statements (ii) and (iii), again, show how the entailment relation is unbounded. Our view of the world is surely bounded.

This boundedness has consequences for understanding knowledge for non-ideal agents. When an agent knows that she has hands, we may view her as knowing that "she is in *a* hand state", with an emphasis on the indefiniteness of what she knows. I want to highlight the indefinite article 'a' in 'she is in a hand state'. As from (ii) and (iii), there are countless ways of being in a state with hands or being in a state without hands, but, clearly, non-ideal agents cannot entertain or consider as possible every one of these ways. What non-ideal agents know is restricted to the domain of which they're aware.

[47] As well, "I have a hand" does not entail that "I am not wrong about having hands", because one may have hands and think that one does not, and be wrong.

One way to understand the skeptical dialectic is that skeptics are able to take advantage of our unawareness, in the strict sense of awareness as presented in section 3.5.1. Before encountering a skeptic, we may take ourselves to know that we have hands. But the skeptic, when successful, suggests a way of being handless that we did not previously consider, a way of which we were unaware. When the skeptic says "do you know whether you are not a handless brain in a vat", she introduces something new, a new state perhaps or a new syntactic element into our vocabulary (the item "brain in a vat"). Such a new syntactic element is, effectively, one more way to fill in statement (iii).

My response to the above skeptical argument is similar to a contextualist response, though I wish to emphasize our cognitive limitations and our bounded view of the world. David Lewis' (1996) contextualist response takes advantage of restricted quantifier domains for the quantifier "every". On his view, to know a proposition is to eliminate *every* alternative to p, and what counts as an alternative is determined by context. My model has similar mechanics. Rather than viewing alternatives as context-dependent, I view the indefinite nature of knowledge as context-dependent. In ordinary contexts, I know I have hands. This is to be paraphrased as "I know I am in *a* hand state" (I am in a state where I have hands). Yet, what I know is indefinite. I know I am in *a* state. I know that the hand state I am in is not one where I am a handless scuba diver, or a handless brain surgeon, but this list is not complete.

Context renders my view of the world comparatively more or less complete. When I was younger, let's say at a time before I ever took a philosophy class or met a philosophical skeptic, I was unaware of brains in vats. I also knew I had hands. However, my view of the world was somewhat incomplete because I had never heard of evil geniuses or Cartesian demons or brains in vats. I had also never heard of Jupiter's moon Io. Yet I eventually learned of these things. One way to understand a successful skeptic is as someone who is able to introduce a new element of which

I was previously unaware. If I cannot rule out this new element when the skeptic raises it (such as being a handless brain in a vat), then the skeptic wins (but perhaps only in that context).

Now, even though I wasn't aware of brains in vats and Cartesian demons when I was young, I am aware of them now. However, this doesn't mean that such information was available to my cognitive processes just a few minutes earlier. In a more limited sense, this information was not available for processing. Though I've heard of brains in vats and Cartesian demons, I wasn't considering or thinking about them. Another way of saying this was that this information was not "salient".

The lesson I drew from Stalnaker (1999) is that knowledge is closed under "available information". I interpreted "available information" in terms of a broad notion of awareness and gave an awareness model in section 3.5. Depending on the particular case (and our cognitive abilities), we sometimes know a proposition when we *know* it is entailed by something we know, and we sometimes know a proposition when it is simply entailed by something we know. What matters is whether this information is available to cognition. Viewed in this light, the skeptic aims to augment our "available information". One way of understanding the introduction of skeptical hypotheses is as tantamount to providing new vocabulary and thereby new possibilities that we are not presently considering or that we had never considered.

3.7.2 The Uninformativeness of Deduction and Not Knowing Skeptical Propositions

Return to the argument from the beginning of section 3.7. In light of my remarks above, premise (S1) is not quite correct ("if I know I have a hand, then I know I am not a brain in a vat"). Let 'h' denote the proposition "I know I have a hand" and let 'b' denote "I am a brain in a vat". By the schema illustrated in (ii) and (iii), h entails "$\neg(\neg h \wedge b)$"; or, h entails that it is not the case that I'm not a handless

brain in a vat.[48] Therefore, (S1) needs to be modified to read "if I know I have a hand, then I know I am not a handless brain in a vat".

Following the original skeptical argument, I took it that we do not know we are not brains in vats. Surely this is intuitively correct.[49] But because the consequent of (S1) has changed, we need to ask whether we know that we are not a handless brain in a vat. On the surface, this change seems inconsequential. By a quick reading it also seems intuitive that I don't know I'm not a handless brain in a vat. But consider the logical form of what this claim amounts to: because $h \Rightarrow \neg(\neg h \wedge b)$, the claim is that I don't know $\neg(\neg h \wedge b)$. Yet the proposition "$\neg(\neg h \wedge b)$" is logically equivalent to "$h \vee \neg b$". So, when it is asked whether one knows that one is not a handless brain in a vat, one is effectively being asked whether one knows that one *either* has hands or is not a brain in a vat.

These considerations should act to deflate the rhetorical power of the skeptic. Viewed in this way, one way of understanding the dialectic proceeds as follows. We begin by taking ourselves to know that we have hands. The skeptic points out that this piece of knowledge entails that we are not a handless brain in a vat (which is supposed to sound too demanding for our epistemic situation). But what the skeptic has really pointed out is that having hands entails that either we have hands or we're not a brain in a vat. And it seems much easier to stay steadfast and claim that we know *this*: if I know I have hands then I know I have hands or I'm not a brain in a vat, by the semantic version of the rule from propositional logic "$\vee$-introduction" (that is, we can infer this from the meaning of "$\vee$").

The theoretical principle behind this kind of response concerns the "uninformativeness of deduction". By the nature of entailment, if p entails q, then "the information in q is already contained in p". Which is to say, if p entails q, then

[48] Precisely, $h \Rightarrow \neg(\neg h \wedge b)$.

[49] That is, I mean that it is intuitive to respond, when asked, that one does not know one is not a brain in a vat. This is not a claim about what we know in ordinary contexts.

every way the world might be when it is a p world is also a q world. At any rate, the explanation is clear concerning the rule $\vee$-introduction. The proposition p entails the proposition $p \vee q$ because, in essence, $p \vee q$ says less about the world "p-wise" than p alone.

The uninformativeness of deductive consequence makes the skeptic's argumentative strategy look epistemically odd. What she asks of her interlocutor is to consider a logically weaker proposition than was claimed to be known. If I claim to know h, she asks if I know the weaker $h \vee \neg b$. But shouldn't I admit that I know this disjunction? After all, I know the left disjunct! Viewed in this way, the proposition b is really a *non sequitur* (on its own).

Now, the skeptic may protest that I've mishandled the structure of the dialectic. First, the skeptic may suggest that what she's done is to provide a context whereby it is rational to suspend judgment as to whether h, then present an epistemically demanding alternative to h that I do not know. Viewed in this way, the dialectic goes as follows. We begin by taking ourselves to know that we have hands. The skeptic wishes to challenge us on this, and asks that we suspend judgment with respect to h (we get to keep our reasons for h, we must only reconsider anew whether h). The skeptic next presents an alternative "$h \vee \neg b$" which is incompatible with h, and asks how we can know *this*.

Again, there is something epistemically odd about the dialectic. Surely, if I am not allowed to claim h for sake of argument, then I no longer have reason for $h \vee \neg b$—but this has little to do with the content of b. For, *any* proposition r such that I lack reason for could serve this role (that is, ordinary and not just skeptical propositions could serve this role). If 'r' denotes the proposition "Meade wore white socks at Gettysburg", then the skeptic suggests that I have no reason to endorse $h \vee r$ because I am to suspend judgment on h (because I know nothing about civil war sock color). But this still does not seem correct. Though I suspend on h, I get to keep my reasons for h, which are reasons for $h \vee r$ (again, for any r). Of course,

the skeptic wishes to pick some r which makes my reasons for h seem problematic. My reasons for r presumably include perception or perceptual experiences as of having hands. The clever skeptic picks an alternative r such as "I am a brain in a vat" that purport to cause doubt about my reasons for h (i.e., if I am a brain in a vat then my experiences do not seem to provide good reasons for thinking I have hands). But such a move muddies the rules *and* the force of the dialectic—if I can keep my reasons for h, even while suspending judgment about h, then I have reason for $h \vee r$, by simple logic. If I cannot keep my reasons for h, then *of course* I have no reason for $h \vee r$ when I have no reason for r, skeptical or otherwise. But this isn't surprising. If the skeptic forces us to begin the dialectic with an impoverished epistemic position then it is clear we will not emerge as knowing much.

However, the skeptic might instead suggest that skeptical alternatives work to attack my original reasons or evidence for h. I claim to know that I have hands because I have perceptual evidence that supports the view that I have hands. But the un-eliminated possibility of brains in vats seems to undercut the evidential support such perceptual experience provides. For, if I *were* a brain in a vat, my perceptual experience would not provide appropriate evidence for thinking that I have hands because my perceptual experience would not be anchored to or correlated with or about the world in the correct way.[50] Surely this is an avenue for the skeptic to pursue. But this particular response by the skeptic does not directly concern the issue of closure. Because it is not my task to refute all skeptical arguments, this skeptical argument is beyond the scope of my present work.

Two final remarks. My previous response above to the skeptic feels similar in spirit to that of Moore (1962). This response may be thought of as dogmatic, in that I don't feel the skeptic is entitled to my giving up reasons for the proposition "I have hands" when she suggests that I do not know what is entailed by "I have hands".

[50]There are various ways one might suggest that brain in a vat experiences do not confer evidence or reasons or justification for external world beliefs.

In this sense I'm not saying much that is new. But, second, I wanted to make clear the oddness of the skeptic's strategy in employing deductive consequence. Putting stress on the fact that "deduction doesn't tell us something new", it is clear that our reasons for a proposition are *ipso facto* reasons for any logically weaker proposition. My reasons for h are reasons for $h \vee b$, for any b. So it is epistemically odd for a skeptic to try such a strategy.

Yet, it must be said that the platitude "deduction is uninformative" is only half true (or true under one particular interpretation). Certainly, mathematicians gain new knowledge. Indeed, this is how I began this chapter. But perhaps the best thing to say, something consonant with the awareness models presented above (as a sentential model), is that mathematicians learn new "sentential information". She knew, in advance, that what she would prove would be *a* tautology (it would be true in every possible world and, hence, would not inform her of which possible world is the actual world). But she didn't know the particular *sentence* that described the tautology.

I do not want to discount the rhetorical power of many skeptical arguments. Without a way to parse the logical form of entailed skeptical alternatives, propositions such as "I am not a handless brain in a vat" may seem beyond our cognitive ken. We can't know *that*, it often seems. And if we don't think we can know *that* and we recognize the logical entailment, then we should recant the ordinary proposition. Yet, it can be easy to "see" the logical connection between h and $h \vee \neg b$ (and, hence, that our reasons for h are reasons for $h \vee \neg b$) after we take our first logic class or when it is pointed out what is meant by "$\vee$". But surely there are other logical connections that *are* difficult to "see", such that we know there is some kind of evidential connection between the propositions but we cannot identify that there is an entailment relation. When the skeptic is able to find these "difficult to see" connections, the skeptic is able to drive a wedge between our acknowledgment of an entailment, and our seeing that the reasons we have for what entails, are really

reasons for what is entailed.

3.8 Conclusion

The search for true closure principles helps make sense of the platitude that deduction is a way of extending our knowledge. But another equally important observation is that our cognitive limitations impose limitations on our deductive powers. The problem of logical omniscience helps makes this clear because it focuses attention to the nature of our cognitive limits. Previous attempts to fix extant closure principles with further and more complicated antecedent conditions, in effect, aim to strengthen the agent's epistemic position so as to overcome such cognitive limitations. For instance, "competent deduction" is stronger than mere entailment alone because the agent is somehow applying the proper rule for deduction. One could strengthen the antecedent condition further by adding conditions that the agent has perfect memory, no syntactic processing mistakes, etc. Perhaps this would yield a (trivially) true closure principle, but it also would render the agent nearly logically omniscient.

At any rate, such attempts at fixing closure principles piecemeal, I argued, don't show the real tradeoff between our deductive abilities and our inherent cognitive limitations. I offered a model of information access and awareness that helps clarify this tradeoff. Admittedly, this model is still somewhat idealized. Yet it makes clear high-level or structural properties of information access and our ability to extend what we know by deduction.

CHAPTER 4

COGNITIVE LIMITATIONS AND KNOWLEDGE

4.1 Introduction

One bar against knowledge comes from the threat of skepticism. The general skeptical strategy is to try and raise the standards for the achievement of knowledge to such a level as to be unattainable. Yet another bar to our knowledge comes from our cognitive limitations. It is fairly mundane to note that *particular occasions* of cognitive disfunction rob us from knowing. When I fail to remember p at t, and have no other evidence for p, I might be properly judged as not knowing p. For instance, when I'm in the airport and forget my flight number (because I'm in such a rush) it seems appropriate to judge that I don't know that my flight number is 2713. When my vision is poor, say, from too much eye strain, I may fail to know some external world proposition ("Julie is standing outside my office").

These *particular* occasions of cognitive disfunction (poor memory, poor vision, etc.) rob us of knowing *particular* propositions, and we make these judgements ("Billy doesn't know p") on an intuitive and typically case by case basis. Again, that these particular lapses in our cognitive function destroy knowledge is somewhat mundane. But a case can be made to show how aspects of *general* cognitive disfunction or limitation rob us from knowing in a systematic and correspondingly *general* way. These arguments will not purport to show that we never know, hence these arguments are not full-blown skeptical arguments. Yet, they will establish systematic bounds on what we can know *in principle*, and they will explain why various cognitive limitations provide boundaries on our knowledge, and what these boundaries are.

Timothy Williamson (2000) has argued that one can know without being in a position to know that one knows.[1] His argument tacitly relies on the premise that our powers of perceptual discrimination have limits, yet he does not fully exploit this fact. In what is to follow, I will set up Williamson's argument and then show how one may argue for his conclusion in a more simple and general manner. This new argument does not rely on two controversial premises that Williamson employs, one premise of which Assaf Sharon and Levi Spectre (2008) have recently challenged. I will assess this challenge and show how my new argument is immune from their challenge.

Next, I will explore other systematic cognitive failures (including various memory and information processing errors) and show how they provide limits for what we can know in principle. I will argue that limitations in perceptual discrimination are not the only reason to doubt that we always possess higher-order, or iterated knowledge. For many occasions, the structure of memory, information processing errors, and general unawareness rob us of higher-order knowledge. Such cognitive phenomena help show that we can know a proposition, but fail to know that we know, or that we can not know a proposition, and fail to know that we don't know this.

I conclude by offering an argument that helps tie together the insight provided by Williamson's argument, my simpler argument, and Sharon and Spectre's challenge. At the highest level of abstraction, always having higher-order knowledge of one's knowledge concerns the transitivity of epistemic possibility. I will argue that the transitivity of epistemic possibility is not epistemically benign—such a restriction entails the existence of what are called "self-evident events". If knowledge is self-evident, then we have deep self-knowledge of the structure of our epistemic possibility function. But we do not have such deep self-knowledge, so, I argue, we do not have transitive epistemic possibility functions. Such a result shows how we

[1] Williamson (2000), p. 114.

may know a proposition but fail to be in a position to know that we know this proposition.

4.2 Williamson on Margins and Iterations

Williamson (2000) considers the case of Mr. Magoo, someone not too unlike each of us in his cognitive abilities. Mr. Magoo considers a tree in the distance and wonders about its height. Now, Mr. Magoo has decent vision, but the tree in question is somewhat far away. His vision is such that he cannot tell to the nearest inch how tall the tree is just by looking. So, supposing he has no other sources of information at hand, Mr. Magoo doesn't know how tall the tree is to the nearest inch.

However, by looking, Mr. Magoo does know that the tree isn't 60 inches tall and he also knows that the tree isn't 6,000 inches tall. Suppose that the tree is 600 inches tall, but Mr. Magoo does not know that it is. Importantly, Williamson stipulates of Mr. Magoo:

> even if he so judges [the tree to be i inches tall] and in fact it is i inches tall, he is merely guessing; for all he knows it is really $i-1$ or $i+1$ inches tall... Anyone who can tell by looking that the tree is not i inches tall, when in fact it is $i+1$ inches tall, has much better eyesight and a much greater ability to judge heights that Mr. Magoo has.[2]

From these remarks, Williamson suggests the following is true:

> (1) Mr. Magoo knows that if the tree is $i+1$ inches tall, then he does not know that the tree is not i inches tall.

Though (1) is not uncontroversial (I discuss criticism of this proposition in the following section), the thought is that (1) follows from the description of Mr. Magoo, someone with poor eyesight who knows he has poor eyesight.

[2]Williamson (2000), p. 115.

Next, in order to show that one can know without thereby being in a position to know that one knows, he assumes a version of the KK principle to derive a contradiction. Assume it is the case that for every entertained and known proposition p, that Mr. Magoo knows that he knows p. One version of this statement, initially given by Sorensen (1988), is the following:

> (KK) For any pertinent proposition p, if Mr. Magoo knows p then he knows that he knows p.[3]

Further, assume that Mr. Magoo is attentive and has reflected on his knowledge and lack of knowledge concerning the tree, his poor eyesight, and the possible heights of the tree. Let's stipulate that Mr. Magoo is competent at deduction. That is, let's say that the following is also true of Mr. Magoo:

> (C) If p and all members of the set X are pertinent propositions, p is a logical consequence of X, and Mr. Magoo knows each member of X, then he knows p.

These somewhat minimal facts about Mr. Magoo are enough to derive a contradiction. For, we stipulated that Mr. Magoo knows the tree is not 60 inches tall (vision, at least, apprises him of this fact). But by (1), (KK) and (C), Mr. Magoo then knows that the tree is not 61 inches tall. To see this, note that by (KK) Mr. Magoo knows that he knows the tree is not 60 inches tall. By (1) and (C), Mr. Magoo knows that the tree is not 61 inches tall. The reasoning continues. Again, by (1), (KK) and (C), Mr. Magoo then knows that the tree is not 62 inches tall. Clearly, Mr. Magoo can continue reasoning this way until he finally concludes:

> (2) Mr. Magoo knows that the tree is not 600 inches tall.

[3]See Williamson (2000), p. 115, for a discussion of this principle and Sorensen's contribution.

As stipulated, (2) is false. So, one of the premises used in the argument must be rejected, as the reasoning is valid. Williamson suggests that each individual step of reasoning follows deductively, so they are not suspect. Also, it is clear that Mr. Magoo knows the tree is not 60 inches tall (we could easily say instead that the tree is not 1 inch tall and similarly derive the contradiction). So that leaves either (KK) or (C) to be rejected.

First, consider (C). While there is disagreement about the tenability of closure principles on knowledge, (C) is supposed to represent a general closure principle. The intuitive idea behind (C) is that deduction is a way of extending one's knowledge. All that is really needed is something like

> (C′) Knowing each of $p_1, p_2, \ldots, p_n$, competently deducing q, and thereby coming to believe q is a general way of coming to know q.

And (C′) is very plausible. Nonetheless, Williamson argues in favor of (C) and rejects (KK). Hence, we see that by way of Mr. Magoo's poor perceptual discrimination, he cannot always know that he knows.

However, Williamson is clear that he intends to argue against (KK) *and* the following weaker principle:

> (KK_{PTK}) If S knows p, then S is in a position to know that S knows p.

To understand (KK_{PTK}) it is necessary to explain what it is to be "in a position to know" a proposition. The idea of being in a position to know a proposition is supposed to be familiar. Williamson says,

> To be in a position to know p, it is neither necessary to know p nor sufficient to be physically and psychologically capable of knowing p. No obstacle must block one's path to knowing p. If one is in a position to know p, and one has done what one is in a position to do to decide

> whether p is true, then one does know p. The fact is open to one's view, unhidden, even if one does not yet see it.[4]

As examples, it is plausible to think that I am in a position to know whether 1143 is prime. Were I to exercise my competence in division and finding factors of numbers, I would know whether 1143 is prime. In a manner of speaking, this proposition ("1143 is prime") is within my ken.[5] Yet, I am not now in a position to know how many grains of sand are currently on Ocean Beach in San Diego. I've done what I'm in a position to do to resolve whether this is true but I still do not know, so I'm not in a position to know this. The number of grains of sand on Ocean Beach is beyond my epistemic ken (given my evidence).

Williamson then argues that Mr. Magoo is not always in a position to know that he knows, when he knows. For, by stipulation, Mr. Magoo has considered whether he knows p, for every proposition p pertinent to the argument. Because Mr. Magoo does not always know that he knows, and has attentively considered each proposition p pertinent to the argument, he is not in a position to know that he knows.

4.3 Comments on Williamson's Argument

Given his previous work, one might suspect that Williamson's argument proceeds by taking advantage of the vagueness of the term 'know'. After all, the progression of Mr. Magoo's reasoning looks similar to a sorites series (Mr. Magoo knows the tree isn't 61 inches tall, then Mr. Magoo knows the tree isn't 62 inches tall...). Williamson disputes this, and claims that "the crucial point is that the premises of the argument are not justified by vagueness in 'know' but by limits on Mr. Magoo's

[4] Williamson (2000), p. 94.

[5] Interestingly, the English 'ken' is related to the Dutch and German 'kennen', from an Indo-European root shared by 'can' and 'know'. Broadly, to be in a position to know is *to be able to* (can) know.

eyesight and his knowledge of them".[6]

In the following section I will show that Williamson's conclusion is actually easier to reach than his argument may suggest. In particular, I will show that Williamson does not need to assume closure nor does he need proposition (1)—these are then to be viewed as inessential premises in his argument.

However, Sharon and Spectre (2008) have criticized Williamson's argument by focusing on the acceptability and applicability of proposition (1), so it is worthwhile to consider the claim in more detail. As I mentioned above, Williamson suggests that this proposition is true of Mr. Magoo, and, presumably, it is true in virtue of the description of Mr. Magoo. As Sharon and Spectre (2008) say, "the reason to think that (1) is true is that, knowing the limitations of his own visual discriminatory abilities, Mr. Magoo knows that if the tree is $i+1$ inches tall he cannot observe that it is not i inches tall".[7] What are the merits of proposition (1)?

First, I want to point out that (1) assumes that Mr. Magoo has detailed knowledge of his perceptual discriminatory abilities. What he knows about his perceptual faculties is specified to the nearest inch. I mention this because there is the worry that a proposition as specific as (1) is unrealistic for typical human subjects. I know my perceptual faculties do not feature *perfect* discriminatory capabilities, at least because I know they've been wrong in the past. But what we typically know about the limits of perception is somewhat vague. I know that if the saguaro cactus outside my window is eight feet tall, then I don't know that the cactus is not somewhat

[6]Williamson (2000), p. 118.

[7]Sharon and Spectre (2008), p. 290. Note well that in their formulation of the problem, Sharon and Spectre give (1) as (WP): Mr. Magoo knows that (if he knows the tree is not i inches tall, then the tree is not $i+1$ inches tall). They call (WP) an "intuitive formulation" of (1), and they suggest that (WP) is an acceptable contraposition of (1), given that Mr. Magoo has deduced everything that follows from his knowledge and on the assumption of closure and the factivity of knowledge. Because I will argue that closure is inessential to Williamson's argument, I will focus on (1) rather than (WP).

shorter or taller (than eight feet), where "somewhat shorter" and "somewhat taller" are vague predicates.

The question is whether this matters for a discussion of the KK principle. For now, it matters because Williamson employs (1) in his argument and takes advantage of the specificity of the proposition. It is unclear whether one could recreate Williamson's chain of reasoning using a vague restatement of (1). However, as I mentioned above, I will argue in the next section that proposition (1) is not essential for a version of Williamson's argument. Yet there are additional reasons to worry about proposition (1).

Sharon and Spectre (2008) offer two kinds of criticism against (1) that, they claim, blocks Williamson's denial of (KK). Their first criticism concerns whether (1) is too broad. They show that if (1) is true of Mr. Magoo in *any* situation then Mr. Magoo will never know the height of the tree. But, intuitively, Mr. Magoo *could* gain knowledge about the height of the tree from sources other than his perceptual faculties. Perhaps Mr. Magoo takes the time to measure the tree or is able to ask an expert arboriculturist working with the tree. As Sharon and Spectre (2008) say,

> The point is that the proposition is applicable only to cases in which Mr. Magoo knows he has no other sources of knowledge and that the ones he does have do not allow him to know the precise height of the tree".[8]
> ...(1) does not pertain to deductive knowledge, nor to knowledge by measurement, reliable procedure, self-reflection, clairvoyance, testimony, and so forth.[9]

The lesson from these observations, then, is that it seems there should be some modification to (1), restricting its range of applicability.

[8]Sharon and Spectre (2008), p. 293.

[9]Sharon and Spectre (2008), p. 294. Note, again, that Sharon and Spectre discuss (WP) which they claim is equivalent to (1). I gave the above quote referencing (1) rather than as they do, referencing (WP), for sake of continuity.

Now, Williamson would admit that (1) is plausibly true when modeled on Mr. Magoo's perceptual faculties *alone*. In line with this, Sharon and Spectre (2008) consider whether the modification to (1) below could complete Williamson's argument:

> (1L) Mr. Magoo knows that, if the tree is $i + 1$ inches tall, then he does not know *just by looking* at the tree that it is not i inches tall.[10]

As should be clear, (1L) differs from (1) by the additional clause "just by looking". The idea behind the move to proposition (1L) is fidelity to Williamson's original description of Mr. Magoo. It is Mr. Magoo's ability at perceptual discrimination, not abilities in measurement or memory, that bar him from knowing various propositions.

Sharon and Spectre (2008) show that (1L), in conjunction with facts about Mr. Magoo, (C) and (KK) *does not* yield a contradiction. We may still suppose that Mr. Magoo knows that the tree is not 60 inches tall by looking. By (1L), (KK) and (C), Mr. Magoo can *infer* that the tree is not 61 inches tall. But because this new knowledge is the product of inference, not "just by looking", Mr. Magoo *cannot* use this new knowledge in further iterations of (1L), (KK), and (C). This is, of course, because the consequent of the embedded conditional in (1L) contains the clause "just by looking". Hence, Mr. Magoo cannot engage in a chain of reasoning concluding with the false proposition "Mr. Magoo knows that the tree is not 600 inches". Proposition (1L), then, appears to block the *reductio* argument.

Next, I will argue that Williamson has the resources to respond to Sharon and Spectre. In what is to follow I will show, directly, why it is Mr. Magoo's perceptual shortfall that explains why he can know something without knowing that he knows, and not issues about proposition (1) or epistemic closure.

[10]Sharon and Spectre (2008), p. 295.

4.4 An Alternative Argument

The kind of perceptual shortfall Williamson describes of Mr. Magoo concerns his powers of perceptual discrimination. When the tree Mr. Magoo observes is actually i inches tall, due to his poor vision, "for all he knows it is really $i-1$ or $i+1$ inches tall".[11] This is the key feature of Williamson's example. The expression "for all he knows" speaks to what is epistemically possible for Mr. Magoo. Given his discriminatory ability, the tree's being i inches tall is indistinguishable from its being $i+1$ inches tall. Perceptually, the tree's being i inches tall *seems*, *looks like*, and *appears just as* the tree's being $i+1$ inches tall. Similarly, invoking an epistemic sense of "looking", when the tree is i inches tall, it (epistemically) looks just as it would look were it $i+1$ inches tall.[12]

There is a straightforward way to model Mr. Magoo's perceptual limitation, one that shows why Mr. Magoo does not always know that he knows, when he knows. However, to do so will require the resources of a set-theoretic model of knowledge.

4.4.1 Information Structures

In 1976 Robert Aumann, an economist, published a paper that described and characterized properties of knowledge in terms of events (given in set-theoretic terms), rather than in terms of syntactical formulas.[13] The kind and formulation of events Aumann had in mind were those that are typically investigated in probability and statistics. Unbeknownst to him, his formulation of knowledge turned out to be equivalent to traditional Kripke structures, the common epistemic logic that bears a relation to modal logic.[14] The formalism is simple, and it will be seen to offer a

[11] Williamson (2000), p. 115.

[12] The epistemic sense of "looking" conveys information about reasoning, as when we say "it looks like rain" when it is not presently raining.

[13] Aumann (1976), "Agreeing to Disagree".

[14] See Fagin et al. (1995). Technically, information structures (also called Aumann structures) are equivalent to Kripke frames.

flexible and expressive tool for describing cognitive limitations.

An "information structure" is a pair (Ω, P), with a set Ω of states and a "possibility function" P that accounts for what the agent considers as "epistemically possible" at a state ω.[15] Elements $\omega \in \Omega$, the states of Ω, are "full descriptions of the world". P is a function that assigns, for each $\omega \in \Omega$, the set of states that are epistemically possible for agent S at ω. The idea is that $P(\omega)$, the set of possibilities at ω, denotes the set of states that the agent cannot distinguish between.[16] When $\omega, \omega' \in P(\omega)$, then for all S knows, ω seems or appears (epistemically) just as ω' does (at ω).

Because information structures do not have a *language* (as epistemic logics do), what is known to agents are "events", where an event is a set of states.[17] Knowledge is then defined as follows:

> **Set-Theoretic Definition of Knowledge**: Let (Ω, P) be an information structure. The event E is known at state ω if and only if $P(\omega) \subseteq E$.

The idea is similar to Hintikka (1962): an event is known when its alternatives are excluded, or, alternatively, S knows E when no non-E state is possible.

4.4.2 Not Knowing That You Know

The above formalism of information structures is enough to model Mr. Magoo's perceptual shortcomings and show why he does not always know that he knows. For all Mr. Magoo knows, when the tree is i inches tall, it is really $i-1$, i, or $i+1$

[15] I gave an introduction to the set-theoretic model of knowledge in Chapter 1.

[16] P is analogous to the accessibility relation in modal and epistemic logics.

[17] Again, events are understood here as they are typically understood in probability and statistics. States are taken as primitive to the model, and events are those of interest to probability theorists. Modeling a six-sided die roll, let the set of states be $\Omega = \{1, 2, \ldots, 6\}$. The event that a die roll is even, then, would be $E = \{2, 4, 6\}$. We say that E obtains or that E is true if the actual state ω is such that $\omega \in E$.

inches tall. Which is to say, when the state is i, Mr. Magoo considers as possible that the state is either $i-1$, i, or $i+1$. So, for Mr. Magoo, his set of possibilities at i is $P(i) = \{i-1, i, i+1\}$.

Consider a perfectly analogous situation. Suppose Mr. Magoo observes a gauge that takes values from the set $\{0, 1, \ldots, 9\}$. Suppose, like in the case of the tree, he has limited powers of perceptual discrimination. When the gauge reads some value n, he cannot discriminate between its being $n-1, n$, or $n+1$. (Suppose also that $P(0) = \{0, 1\}$, and $P(9) = \{8, 9\}$.) It is easy to show that when this is the case, Mr. Magoo's possibility function does not satisfy the following condition:

(P2): If $\omega' \in P(\omega)$, then $P(\omega') \subseteq P(\omega)$.

For, $3 \in P(2)$ but it is not the case that $P(3) \subseteq P(2)$ because $4 \in P(3)$ but it is not the case that $4 \in P(2)$.

Next, if an agent's possibility function does not satisfy (P2), then the agent does not always know that she knows (when she does). That is, if an agent's possibility function does not satisfy (P2), then she violates the KK principle, given here in set-theoretic terms:

(K2): $K(E) \subseteq K(K(E))$.

This condition says that if an agent knows E, then she knows that she knows E, expressed in the formalism of the set-theoretical model of knowledge.[18] (K2) is equivalent to (KK).

To see why it is the case that if an agent's possibility function does not satisfy (P2), then the agent does not know that she knows, consider the following proof that (K2) entails (P2):

Lemma 4.1: (K2) $\Rightarrow$ (P2).

[18]What (K2) expresses, if true, is that any event of knowing E must be an event of knowing that it is known that E, because of the subset relation.

Proof: Assume $\omega' \in P(\omega)$. Assume $\omega \in P(\omega')$ and show $\omega \in P(\omega)$ (this is to show $P(\omega') \subseteq P(\omega)$). (K2) says, if $P(\omega) \subseteq E$ then $P(\omega) \subseteq K(E)$, for any event E. We know that $P(\omega) \subseteq P(\omega)$ and $P(\omega') \subseteq P(\omega')$, trivially. Also, we know that $P(\omega)$ and $P(\omega')$ are events. Hence, by (K2), $P(\omega) \subseteq K(P(\omega))$ and $P(\omega') \subseteq K(P(\omega'))$. Because $\omega' \in P(\omega)$, it must be that $\omega' \in K(P(\omega))$. Because $\omega \in P(\omega')$, it must be that $\omega \in K(P(\omega'))$. By the set-theoretic definition of knowledge, $P(\omega) \subseteq P(\omega')$ and also that $P(\omega') \subseteq P(\omega)$. So $P(\omega) = P(\omega')$. But because $\omega \in P(\omega')$, it must be that $\omega \in P(\omega)$. ■

Hence, by *modus tollens*, if an agent's possibility function does not satisfy (P2) then she violates (K2). As Williamson describes Mr. Magoo, his possibility function is such that it violates (P2) (I showed this above). So Mr. Magoo does not always know that he knows, when he knows.

It is also easy to show that (P2) uniquely characterizes (K2). That is, (K2) obtains in the model if and only if an agent's possibility function P satisfies (P2).

Lemma 4.2: (K2) $\Leftrightarrow$ (P2).

Proof: From Lemma 4.1, all that needs to be shown is that (P2) entails (K2). Assume that E is known by the agent, or that $\omega \in K(E)$, and that (P2) holds. Let $\omega' \in P(\omega)$. By (P2), $P(\omega') \subseteq P(\omega)$. Yet, because $\omega \in K(E)$, it follows from the set-theoretic definition of knowledge that $P(\omega) \subseteq E$ and hence it follows that $P(\omega') \subseteq E$. So $\omega' \in K(E)$. This establishes that $P(\omega) \subseteq K(E)$, so $\omega \in K(K(E))$. ■

4.4.3 Remarks on the Argument

My argument does not require a closure principle for knowledge, as Williamson's does. While it is true that knowledge is closed in all information structures, this

fact about the formalism is not used in the proof. However, if it is still uncertain, it is straightforward to give a model of knowledge where closure fails yet preserves the entailment of (P2) by (K2). In Chapter 3 I gave a model of knowledge in "awareness structures" where closure fails. With appropriate modifications (one would need to translate between a set-theoretic approach and a possible-worlds approach) it can be shown that this result still holds.

As well, my argument against the KK principle does not make use of a principle such as (1). As I've set up the model so far, we need not endow Mr. Magoo with any *knowledge* of his perceptual shortcomings, only the perceptual shortcomings themselves.

Hence, it is clear that Williamson makes use of inessential premises in his argument. For the argument to go through, Mr. Magoo need not have knowledge about his perceptual limitations. Further, those that reject closure principles might not embrace Williamson's arguments when they should—I've shown that one need not endorse closure to reject the KK principle. Williamson has not located the foundational issue.

My argument makes it directly clear why Mr. Magoo doesn't always know that he knows, when he does. His epistemic view of the world is captured with his possibility function $P(\omega)$—his deficit in perceptual discrimination is modeled by this function. Because Mr. Magoo's eyesight is such that the world where the tree is i inches seems just as the world where the tree is $i-1$ or $i+1$ inches tall, his possibility function does not satisfy (P2). But because his possibility function does not satisfy (P2), (K2) must be false for Mr. Magoo.

What's so special about (P2)? As it turns out, (P2) provides a condition of transitivity on epistemic possibility. It is well known in modal logic that the non-transitivity of the epistemic possibility relation yields necessary and sufficient conditions for counter-instances to (K2) and the KK principle.[19] What Williamson's

[19]Williamson (2011) notes that the non-transitivity of the epistemic possibility relation yields

example provides is a clear case, one that is epistemically plausible, of the failure of a transitive epistemic possibility relation. The key claim then, as I used it to begin this section, is the expression "for all he (Mr. Magoo) knows, it is really i or $i+1$ inches tall". When this is true, as it is for Mr. Magoo, (P2) will fail and (K2) will correspondingly fail.

The lesson can be generalized. If, for any reason, we lose our ability in discrimination that renders our possibility function to have the structure it does for Mr. Magoo, we too will fail to always know that we know, when we know. Williamson is surely right on this point. The examples extend well beyond trees. Think about how old the bread is. I cannot tell the difference between its being 120 and 121 hours old. Similarly, I cannot tell the difference between our walk's being 38 or 39 minutes long. I also cannot tell the difference between a cup of tea with 343 or 344 grains of sugar. In these cases we have some knowledge, call it "inexact" knowledge (I know I put a rough teaspoon of sugar in the tea), but because I don't know the boundaries to what I know, I cannot be confident that I will know that I know (what I know).[20]

Just as with Mr. Magoo and the tree, perceptual knowledge is often pervaded by failures of the KK principle because perception does not always yield transitive epistemic discrimination.

4.4.4 More About Mr. Magoo

There is more to be said of Mr. Magoo. Given what we know about Mr. Magoo, his possibility function also violates a further principle:

---

necessary and sufficient conditions for the existence of counter-instances to (K2). But what is missing from this subsequent discussion is a clear explanation for why agents like Mr. Magoo violate transitivity.

[20]These examples and the expression "inexact knowledge" are Williamson's. See Williamson (2000), p. 119.

(P3): If $\omega' \in P(\omega)$, then $P(\omega) \subseteq P(\omega')$.

From the above example of the gauge, it is clear that $3 \in P(2)$ but it is not the case that $P(2) \subseteq P(3)$ because $1 \in P(2)$, but it is not the case that $1 \in P(3)$.

If an agent's possibility function does not satisfy (P3), then the agent does not reason in accordance with the "negative introspection" axiom (axiom 5, in many epistemic logics). For a proof, see Lemma 4.3 below. That is, if an agent's possibility function does not satisfy (P3) then it is not the case that if S does not know p, S knows that S does not know p.[21]

This observation provides a minor reason to doubt Williamson's use of premise (1). I will turn to Sharon and Spectre's argument against Williamson's use of (1) in the following subsection.

Because Mr. Magoo's possibility function is structured such as to violate (P3), it is possible for him to not know a proposition and not know that he doesn't know it. Formally, if (P3) is false then the following is also false:

(K3): $\neg K(E) \subseteq K(\neg K(E))$.

This principle is often referred to as the "negative introspection" axiom, and it expresses the idea that agents "know what they don't know".

Lemma 4.3: (K3) $\Rightarrow$ (P3).

[21]It is straightforward to show that (P3) entails (K3). Assume (P3) and assume that $\omega \notin K(E)$. Assume $\omega'' \in P(\omega)$ and show $\omega'' \notin K(E)$ (this is to establish that $P(\omega) \subseteq \neg K(E)$). Because $\omega \notin K(E)$, this means that $P(\omega) \nsubseteq K(E)$. Then either $P(\omega) \cap K(E)$ is empty or it is not. Suppose $P(\omega) \cap K(E)$ is non-empty. Then there exists some $\omega^\star$ such that $\omega^\star \in P(\omega)$ and $\omega^\star \in K(E)$. Because $\omega^\star \in K(E)$, this means that $P(\omega^\star) \subseteq E$. But by (P3), $P(\omega) \subseteq P(\omega^\star)$, so it follows that $P(\omega) \subseteq E$, so $\omega \in K(E)$, which is a contradiction. So $P(\omega) \cap K(E) = \emptyset$. Now, because $\omega'' \in P(\omega)$ it follows that $\omega'' \notin K(E)$, which was to be demonstrated.

Proof: Assume (K3). That is, assume that if $\omega \notin K(E)$ then $P(\omega) \subseteq \neg K(E)$. Assume $\omega' \in P(\omega)$. Show that $P(\omega) \subseteq P(\omega')$. Because $\omega' \in P(\omega)$, it follows that $\omega' \notin -P(\omega)$. So, $P(\omega') \subseteq P(\omega)$ (from the assumption of (K3)). Now, either $\omega \in P(\omega')$ or $\omega \notin P(\omega')$. Assume that $\omega \notin P(\omega')$. Then $P(\omega) \subseteq -P(\omega')$, so $P(\omega') \subseteq -P(\omega')$. But this is impossible. So, $\omega \in P(\omega')$. This means that $\omega \notin -P(\omega')$. But from the assumption of (K3), this means that $P(\omega) \subseteq P(\omega')$, so $P(\omega) = P(\omega')$. This establishes that $P(\omega) \subseteq P(\omega')$.[22] ■

There is a tension between an endorsement of (1) and a denial of (K3). Williamson's argument must embrace this tension yet my argument is immune from this tension. The tension, stated broadly, is this: for (1) to be true of Mr. Magoo, he must be aware of much of what he does not know by perception. He is "aware of his perceptual ignorance". But because (K3) is false, Mr. Magoo is *not* "aware of his ignorance" generally. These two facts about Mr. Magoo are not irreconcilable, there is no logical inconsistency in both obtaining, however we're left with either a confusing or seemingly unrealistic view of Mr. Magoo. Again, if (1) is true but (K3) is false, Mr. Magoo is aware of his perceptual ignorance but unaware of his ignorance generally.

It is better to argue without proposition (1), as I've done.

4.4.5 Responding to Sharon and Spectre

Sharon and Spectre (2008) have shown that Williamson's use of proposition (1) is problematic. My argument for Williamson's conclusion, a denial of the KK principle, does not require proposition (1), so their argument does not address the essential issue in knowing that you know. As I suggested above, the essential issue concerns

[22]Recall that because this is a set-theoretic model, the negation connective functions as the set-theoretic complement operation.

the transitivity of epistemic possibility. My argument shows how Mr. Magoo's perceptual shortcomings are directly related to his failure to know that he knows, when he does.

I want to recast Sharon and Spectre's worries with Williamson's argument. They first argue that proposition (1) is not true of Mr. Magoo generally. They then argue that relativizing proposition (1) to Mr. Magoo's perceptual faculties (by adding the clause "just by looking") does not yield the contradiction sought by Williamson's original argument.[23] They take as a lesson from these observations, the following:

> The problem is more general... This case exemplifies the tendency to forget that (sometimes) what matters is not only whether you know that p but also *how* you know it. The underlying theme of this paper is that the differences between the modes by which knowledge is acquired have more implications than epistemologists tend to recognize... It may also be suggested (in line with our argument) that the validity of (KK) ought to be determined separately for different modes of knowledge.[24]

I emphatically agree with most of this.[25] I take Williamson't example to show that perception, in many cases, yields "inexact knowledge" and hence, results in intransitive epistemic possibility functions. Hence, ordinary perceptual limits on discrimination yield failures of KK. But, as an addendum, it is not only how you know but what you know that matters. For, inexactness of knowledge often has to do with the precision of what is known. As Sharon and Spectre mention several times, Mr. Magoo might have measured the tree (rather than estimated its height by sight).

[23]To be clear, Sharon and Spectre (2008) do not take themselves as arguing for or against the KK principle, only pointing out a difficulty with Williamson's argument. See Sharon and Spectre (2008), p. 299.

[24]Sharon and Spectre (2008), p. 299.

[25]For similar reasons, I agree that Williamson's use of the Mr. Magoo case for the surprise examination paradox bears more scrutiny.

Yet, even if he measured, Mr. Magoo surely could not tell the difference between the tree's being 601.234643 feet and 601.234644 feet tall (when such information outstrips the resolution of the measuring device). If Mr. Magoo gave a proof that p, he surely could not tell the difference between his proving p 1334 seconds ago and his proving p 1335 seconds ago.

In the following section I explore other epistemic scenarios with an aim to model their epistemic consequences. In a way, this is to carry out Sharon and Spectre's challenge. I will look at various modes of knowledge acquisition and determine their epistemic implications.

4.5 Cognitive Limitations and their Implications for Knowledge

The formalism of possibility functions and information structures provides a wonderfully expressive yet exacting tool to describe and explore cognitive limitations and their implications for knowledge. Mr. Magoo provided one case, but there are many others. Our cognitive limitations destroy knowledge in ways that go beyond inaccurate perceptual discrimination.

4.5.1 Unremarkable Events

It is commonplace to recognize that in many cases extreme events are memorable, but the mix is unremarkable. For good or ill, our cognitive systems are such that we sometimes only notice the unusual or atypical, but miss the routine. As an example, we may notice an exceptionally hot or cold day, but fail to record a typical day. We may notice an excellent student or notice a bad student, but fail to notice an average student.

Perhaps, the existence of such unremarkable events helps partially explain Kahneman and Tversky's (1973) regression fallacy. Kahneman and Tversky found that Israeli air force instructors mistakenly thought that punishing poor pilot perfor-

mance was more effective at improving performance than praising good performance.[26] The attributed reason for the mistaken beliefs was that punishment was typically followed by improved performance but praise was typically followed by worse performance. However, Kahneman and Tversky knew that pilot performance should follow a "bell-shaped" distribution, or a Gaussian distribution. Such a distribution has the property that atypical performance in either direction (either atypically good or atypically bad) is most likely followed by performance closer to the mean. So, statistically speaking, excellent flight performance is most likely followed by decreased performance and poor flight performance is most likely followed by improved performance. When Kahneman and Tversky controlled for this effect, often called "the regression fallacy", they found that praise actually positively impacted behavior better than punishment. But why didn't the Israeli air force instructors notice that their praise was effective? Part of the answer requires unremarkable events. Presumably, the air force instructors did not notice that successive performance that centered around the mean was improved by praise. Instead, the air force instructors paid most attention to extreme performance, where "regression to the mean" is operative. The instructors noticed extreme performance but did not properly record typical performance.

The formalism of possibility functions given above provides a tool to model such a cognitive phenomenon. Suppose agent S is such that she notices extreme events but does not notice typical events. Let the set of states be $\Omega = \{1, 2, 3\}$, where $1 =$"red Ferrari", $2 =$"red Ford Truck", and $3 =$"Truck-zilla" (a monster truck). Suppose that when S is in an extreme state she takes notice, but that when S is in a typical state she does not take notice. Her possibility function can then be described as the following: $P(1) = \{1, 2\}$, $P(2) = \{1, 2, 3\}$, $P(3) = \{2, 3\}$. Further, let $E_1 = \{1, 2\}$ be the event "red car" and let $E_2 = \{2, 3\}$ be the event "truck". Given S's possibility function, S notices that she observes a red car when a Ferrari

[26] See also Bishop and Trout (2005), p. 149.

is present (state 1), but does not notice that she observed a red car when a Ford Truck is present (state 2).

From the given information, it is easy to show that S's possibility function does not satisfy (P2) and, hence, violates (K2). So S may know some event but not know that she knows. To see why her possibility function violates (P2), notice that $2 \in P(1)$ but $P(2) \nsubseteq P(1)$. As a concrete example, S knows E_1 at 1, but does not know that she knows at 1. So S is such that she knows, but she doesn't always know that she knows.

The above model suggests that when we fail to notice unremarkable events we thereby lose higher-order knowledge of what we know. Or, said differently, the model suggests that to always know that we know, we must be free from memory failures that ignore unremarkable events, similar to S above. This result is novel and surprising—without the structure of the model, I doubt that it would be clear to see the relationship between memory failures of this kind and failures of higher-order knowledge.

Real human agents are not always free from these memory limitations. We routinely forget typical and mundane events, while noticing the extremes. From the model, then, these observations serve as evidence against (KK). I will return to the larger issue of the status of (KK) and the weaker (KK_{PTK}) principle in section 4.6.

4.5.2 Unclear on the Details

There are situations in which we are unable to fully process available information. As one example, consider a student who does not take praise or criticism well. Suppose that the student is pessimistic when she hears good news, but optimistic when she hears bad news. Suppose that the student receives a verbal report from her teacher concerning her performance on a recent essay. Let the set of states be $\Omega = \{1, 2, 3\}$, where 1 =“excellent performance on the essay”, 2 =“average performance on the essay”, and 3 =“bad performance on the essay”. Further,

suppose the following: when the student has excellent performance, her teacher tells her "excellent" but the student, due to pessimism, thinks "not bad"; when the student has average performance, her teacher tells her "average" but the student, due to limited optimism, thinks "not best"; when the student has bad performance, her teacher tells her "bad" but the student, due to limited optimism, thinks "not best". The student's possibility function can then be described as the following: $P(1) = \{1, 2\}$, $P(2) = \{2, 3\}$, $P(3) = \{2, 3\}$.

From the given information, it is easy to show that the student's possibility function does not satisfy (P2): this is because $2 \in P(1)$ but $P(2) \nsubseteq P(1)$. By Lemma 4.1 and *modus tollens*, (K2) is not satisfied, so the student may know some event but not know that she knows. As a concrete example, S knows the event $E_1 = \{1, 2\}$ at 1, but does not know that she knows E_1 at 1.

The example of the student shows that failure to process available information has implications for higher-order knowledge. Just as it is with perceptual discrimination failures and not taking notice of unremarkable events, failing to fully process information can destroy higher-order knowledge. In the case of the student, when she performs well on the essay she knows that she didn't perform badly. But she doesn't know that she knows this. Intuitively, the student faces a kind of epistemic uncertainty which results from her inability to fully process information. This uncertainty prevents her from appropriately distinguishing between her knowing that she knows she did well and not knowing that she did well.

Pessimism and optimism are not the only sources of information processing failure. Some information may be too unpleasant or uncomfortable for an agent to fully process. Psychologists have identified cases where a spouse or close family member is unable to fully process the heinous action of a related individual. Such relatives may only understand that "something is wrong" without understanding the details. If their possibility function resembles that of the student's, the relative may know that something is wrong, but not know that they know this.

4.5.3 Unawareness

At a general or high-level of abstraction, the cognitive phenomena of confirmation bias and selective attention have in common that the reasoner takes notice of some salient event, but fails to take notice of non-salient events. This is to be distinguished from the case of extreme and typical events. The paradigmatic example of confirmation bias is when an agent recognizes confirming evidence as evidence, but does not recognize disconfirming evidence as evidence. The paradigmatic example of selective attention is when an agent focuses on one kind of event and remains unaware of another kind of event. Though the phenomena of confirmation bias and selective attention are distinct, they share a structural similarity.

Consider a simple model of this cognitive disfunction.[27] Suppose Mr. Magoo observes a gauge that takes values from the set $\{00, 01, \ldots, 99\}$. Further, suppose Mr. Magoo takes notice when the ones digit coincides with the tens digit, but when they do not, Mr. Magoo only notices the ones digit.

For this simple case, Mr. Magoo's possibility function for the state 22 would be $P(22) = \{22\}$, because he takes notice of the coincidence of digits, but his possibility function for the state 23 would be $P(23) = \{03, 13, 23, \ldots 93\}$. Mr. Magoo's possibility function is similar for all other cases where the digits match and do not match.

It is easy to show that Mr. Magoo's possibility function violates condition (P3) but satisfies (P2). To see that it violates (P3), it is true that $33 \in P(23)$, but $23 \in P(23)$ and it is not the case that $23 \in P(33)$. To see that the possibility function satisfies (P2), we must consider the "same digits" case and its complement. In the "same digits" case, (P2) is trivially satisfied because Mr. Magoo's possibility function returns a singleton state. For the complement, suppose $w' \in P(w)$. If w' has the "same digits" attribute, then $P(w') \subseteq P(w)$ because $P(w')$ returns a singleton state. If w' does not have the "same digits" attribute then $P(w') = P(w)$,

[27] Example adapted from Rubinstein (1998).

so trivially $P(w') \subseteq P(w)$.

What can be learned from such a model? As I've shown above, when an agent's possibility function violates (P3), she thereby violates (K3). Which is to say, when an agent systematically fails to take notice of a salient property such as the "same digits" attribute, it is then possible for her to not know that she does not know some event. In other words, phenomena like confirmation bias and selective attention make it possible for an agent to become unaware of their ignorance.

A further interesting consequence of the model is that the agent will know that she knows, when she does. Which is to say, when an agent meets the structural conditions of model presented above, the phenomena of confirmation bias and selective attention will not block her from knowing that she knows. This is a remarkable implication from the model. Confirmation bias and selective attention are thought to be non-ideal (wouldn't it be epistemically better to be aware of every event when it happened?), but the model shows the *particular way* in which these phenomena negatively impact our reasoning.

Of course, one explanation for why confirmation bias is epistemically non-ideal or irrational is that those agents subject to confirmation bias do not take full advantage of their evidence. There is a sense in which they are guilty of violating a total evidence requirement. But what this model shows is that confirmation bias (when it meets the structural requirement given above) also has implications for higher-order knowledge. When agents become systematically unaware of events as they happen, they may thereby fail to appreciate what they do not know—they may not know and also not know that they do not know.

4.5.4 Missing Ambiguities

In the previous subsection, Mr. Magoo didn't take advantage of information that, in some sense, he should have taken advantage of. In this way his possibility function was more coarse than it optimally should have been. But sometimes agents

mistakenly miss complexities in their information and ignore ambiguity.

Suppose Mr. Magoo is overconfident in his ability to draw implications from speech, and doesn't recognize ambiguities when they are present. Suppose Mr. Magoo's communicative partner tells him:

(A) The girl hit the boy with the cap.

Suppose that Mr. Magoo initially considers four states of the world: (i) the girl used the cap to hit the boy (GCB), (ii) the boy used the cap to hit the girl (BCG), (iii) the girl hit the boy who was wearing the cap (GBC), and (iv) the boy hit the girl who was wearing the cap (BGC).

Because Mr. Magoo is not sensitive to ambiguity, when he learns (A), suppose his possibility function has the following structure: $P(GCB) = \{GCB\}$, $P(GBC) = \{GCB\}$.[28] That is, Mr. Magoo does not see that the information in (A) alone (independent of context or other informational cues) does not exclude state GBC. Optimally, Mr. Magoo's possibility function should have the form: $P(GCB) = \{GCB, GBC\}$, $P(GBC) = \{GCB, GBC\}$.

What are the implications for Mr. Magoo when he reasons in the sub-optimal way? When Mr. Magoo misses the ambiguity in statement (A) he potentially excludes the actual state, which he would not do if his possibility function had the optimal and more coarse structure. For instance, when Mr. Magoo is told (A) but it is the case that the girl hit the boy wearing the cap, Mr. Magoo only considers GCB as possible. As such, Mr. Magoo's possibility function violates the following principle:

(P1) $w \in P(w)$.

[28]For simplicity I'm ignoring Mr. Magoo's possibility function for when the state is BCG and BGC and Mr. Magoo is told (A).

This principle says that an agent never excludes the actual state from the set of epistemically possible states. When (P1) is violated, the following principle of knowledge is also violated:

(K1) $K(E) \subseteq E$.[29]

This principle is the familiar principle of the factivity of knowledge. According to (K1), if S knows E, then E must be the case (E must contain the actual state w, or $w \in E$). Alternatively, in the language of epistemic logic, we can say that when Mr. Magoo's possibility function violates (P1) his "accessibility relation" is not reflexive.

Interestingly, we can see that from the structure of Mr. Magoo's possibility function, his failing to recognize ambiguity does not bar him from knowing that he knows or that he knows what he does not know. For instance, when the state is GCB, according to the model, Mr. Magoo knows GCB.[30] But Mr. Magoo also knows that he knows, for his possibility function satisfies (P2). Likewise for (K3). Because Mr. Magoo's possibility function satisfies (P3), he also has knowledge of his ignorance. Hence, we can see that missing ambiguities in language (potentially) only serves to allow us to exclude the actual state, but does not destroy our higher-order knowledge that we know. Such a result would be hard to come by without the formalism that information structures afford.

The case of missing ambiguities is related to an extreme cognitive disfunction. From above, we can see that an agent who systematically fails to see ambiguity might exclude the actual state from being possible. Call an agent who makes such

[29]For further discussion, see Chapter 1.

[30]My intuition is not firm as to whether Mr. Magoo really knows GCB. Against the model, it seems as if Mr. Magoo is lucky—for, as all he really knows, it is possible that GBC. On the contrary, much of language has hidden ambiguity that can be uncovered with dedicated effort. That is, most things we are told have ambiguous readings if properly scrutinized. We learn things from testimony because context often resolves any ambiguity. But because the case I've given abstracts away from issues of context, it is hard to have a clear intuition about the case.

an error "deluded": an agent S is deluded when she excludes the actual state. Now consider an extreme form of cognitive delusion.

Suppose Mr. Magoo is disposed to automatically resolve all ambiguities in language and reasoning. That is, suppose he is so constituted so as to reduce his possibility function $P(\omega)$ to some singleton state ω', at every $\omega \in \Omega$. Which is to say, in this case, Mr. Magoo takes his information to resolve any and all potential defeaters and relevant alternatives, for any such information or evidence he possesses. One way he could do this, though terribly irrational, would be to reduce his possibility function randomly, so that $P(\omega) = \omega^\star$ for a randomly picked $\omega^\star$. In this case, Mr. Magoo would be deluded, and mostly likely not satisfy either (K2) or (K3).

A slightly less extreme form of the previous case would be a dogmatic agent, one who reduces her possibility function to the same singleton ω' for every set $P(\omega)$. The psychological interpretation of such an agent is someone who views the world in one fixed way, regardless of the evidence. Such an agent would be deluded, but, surprisingly, such an agent would not violate principles (P2) or (P3). An agent who is dogmatic in this way would know that she knows (when she knows) and have knowledge of her ignorance.

4.5.5 Too Much Information

It is well known that humans have a finite capacity for processing and retaining information. The question is whether this cognitive limitation, understood in a broad sense, places systematic bounds on our knowledge.

One important aspect of this phenomenon concerns vision. Our visual system cannot handle the amount of information available in the distal environment. Simply, there is too much information available to be processed than can be processed. Let's assume Mr. Magoo is like us in this respect.

The situation before Mr. Magoo is familiar. The characteristic of the phe-

nomenon is that we (often) only have a partial glimpse of the world. A police officer may try to observe the license place of a suspect but only catch the first two letters. The witness may only observe the color of the car but not the particular make of the car.

The intuition is that such cognitive limitations do not impose systematic failures in our higher-order knowledge. Of course, not having more information entails that we will lack knowledge of particular propositions dependent on this unreachable information. But this does not seem to entail that we cannot know that we know (when we know) or that we cannot know that we do not know (when we do not know). I will use the formalism of information structures to vindicate this intuition.

Suppose, again, Mr. Magoo observes a gauge that can take nine-digit values from 000,000,000–999,999,999. Suppose that, due to the natural informational limits imposed on his visual system, Mr. Magoo is only able to observe the first two digits. What implications for Mr. Magoo's higher-order knowledge does such a deficit impose?

To model his possibility function, consider a structurally similar case where values of the gauge range from 00–99 and Mr. Magoo only observes the tens digit.[31] In this case, his possibility function is then $P(\omega) = \{\omega' |$ the tens digit of ω' is identical to that of $\omega\}$. It is straightforward to show that this possibility function satisfies (P1), (P2), and (P3). Hence, Mr. Magoo's knowledge and higher-order knowledge is consistent with (K1), (K2), and (K3). Such informational limits do not directly impose deficits in Mr. Magoo's higher-order knowledge and, hence, do not impose structural limits on his knowledge.

However, given these remarks, it should be clear that Mr. Magoo is somewhat epistemically impoverished, in the sense that he lacks knowledge for propositions that he otherwise would know were his informational limits unbounded. For in-

[31]This reduction in values for the case only makes Mr. Magoo's possibility function easier to describe in a simple notation.

stance, when he cannot process all the available information as from above, he does not know whether the observed number is even or odd. To see this, suppose that the state is 14. Mr. Magoo's possibility function is then $P(14) = \{10, 11, 12, \ldots 19\}$. Consider the event E_e, the event that the true state is an even number. Because it is not the case that $P(14) \subseteq E_e$, Mr. Magoo does not know that the event is even. *Mutatis mutandis* for the case where the event is an odd number.

However, Mr. Magoo does know, when the true state is 14, that the number is less than 21. For, $P(14) \subseteq E_{20}$, where $E_{20} = \{1, 2, 3, \ldots 20\}$. Further, Mr. Magoo knows that he knows this. And he also knows that he knows that he knows this, *ad infinitum*. Because Mr. Magoo satisfies (P3) and hence (K3), he knows that he doesn't know that the true state is even.

So, the intuition that the informational limits imposed on our cognitive architecture do not entail systematic bounds on our knowledge is vindicated. If we had a "finer-grain" picture of the world we would certainly know many more propositions. But that we do not does not entail (by itself) that we cannot know that we know (when we know) or that we cannot know that we do not know (when we do know know).

4.6 Self-Evident Events and Being in a Position to Know That You Know

I've previously argued that using information structures as a model for knowledge helps illuminate the connection between various cognitive limitations and higher-order or iterated knowledge. As a consequence of the model, if agents have perceptual discrimination limitations, do not record unremarkable events in memory, or are unable to fully process information, then these agents do not always know that they know, when they have knowledge. Real, non-ideal human agents often suffer from these cognitive limitations, so real, non-ideal agents don't always know that they know. Yet, the model of knowledge I've presented has the resources to provide for an argument against a related claim, that agents who suffer the above cognitive

limitations are not always even "in a position" to know that they know.

As I mentioned in section 4.2, Williamson (2000) takes as his target not just (KK) but also (KK_{PTK}), the claim that if S knows p, then S is in a position to know that S knows p. Under the familiar understanding of "being in a position to know", (KK) entails (KK_{PTK}) but (KK_{PTK}) does not entail (KK). So far, I've argued against (KK) but I haven't argued against (KK_{PTK}). I turn to this principle next.

Williamson (2000) argued that Mr. Magoo was not always in a position to know that he knows, when he knew some proposition. As was stipulated, Mr. Magoo had considered whether he knew every pertinent proposition to Williamson's argument, Mr. Magoo had specific knowledge (we might call this "self-knowledge") about his own perceptual limitations, and Mr. Magoo was competent at deductive inference. Because Mr. Magoo didn't always know that he knew (Williamson argued (KK) was false), and he had "done what he was in a position to do to decide whether each p was true", Williamson concludes that Mr. Magoo was not always in a position to know that he knows (when he knows).

In what follows, I will also argue that real, non-ideal agents are not always in a position to know that they know (when they know). But I do not want to make assumptions that Williamson makes of Mr. Magoo. Real, non-ideal agents are rarely perfect epistemic analogues of Mr. Magoo—we do not know the precise limitations of our cognitive architecture (we rarely have such precise self-knowledge), nor are we often able to thoroughly consider every pertinent proposition to many important arguments. Instead, I will directly focus on the epistemic requirements of (P2), the restriction on our epistemic possibility function that is necessary for (KK). I will argue that the (P2) places epistemic burdens that real human agents cannot meet and are never in a position to meet. Because real human agents are never in a position to satisfy (P2), we are not always in a position to know that we know.

One way to understand the restriction (P2) is that it provides for a "nested"

possibility function.[32] What (P2) says is that any way $P(\omega)$ obtains, that is for any $\omega' \in P(\omega)$, the possibilities associated with ω' must be contained in $P(\omega)$. Such a nested possibility function provides a kind of epistemic certainty. When events are non-nested (with respect to an agent's possibility function) there are ways for the event to occur *without* the agent knowing that they occur. In general, most empirical and contingent events are such that they *can* happen without us knowing that they happened. For instance, I may be camped out in the basement of the library reviewing old philosophy journals and not notice that it has started to rain. Surely this is a familiar example. In such a case, the event "rain" occurs without me knowing that it is raining, so my possibility function cannot be "nested" inside the event "rain".[33]

(P2), and correspondingly, (K2), entail overly demanding epistemic requirements—requirements that real, non-ideal agents cannot meet. In what is to follow, I will explain these requirements and argue that we (as non-ideal agents) are never in a position to satisfy them.

I will argue that if the KK principle is true (that is if either (P2) or (K2) hold) then knowledge is self-evident. The definition of a self-evident event is the following:

> **Definition**: an event E is self-evident for S if and only if whenever E obtains S knows E. (I.e., E is self-evident for S iff for all $\omega \in E$, $P_S(\omega) \subseteq E$.)

Self-evident events are such that they cannot happen without an agent knowing that they happen. Said differently, a self-evident event E is such that if E obtains then S knows E.

In showing that the KK principle entails that knowledge is self-evident, I will assume that knowledge is factive, so I will assume that (P1) holds. By Lemma 4.2, the KK principle holds if and only if (P2) holds, so I will begin by assuming (P2).

[32] I first described the restriction (P2) in this chapter on page 122. See also Chapter 1.

[33] I will make these claims more precise below.

Theorem 4.1: Suppose (Ω, P) satisfies (P1) and (P2) *and* S knows E. Then there is some self-evident event $E^\star$ that obtains and $E^\star$ entails E.

Proof: Suppose S knows E. This means that $\omega \in K(E)$. Let $E^\star = K(E)$. Because (Ω, P) satisfies (P1) and (P2), (K1) and (K2) hold. By (K1), $E^\star \subseteq E$, so $E^\star$ entails E. Next, show that $E^\star$ obtains and is a self-evident event. By (K2), $K(E) \subseteq K(K(E))$. But because $E^\star = K(E)$, this means that $E^\star = K(E^\star)$, so $E^\star$ is a self-evident event. Because $\omega \in K(E)$, it must be that $\omega \in E^\star$ (by $E^\star = K(E)$), so $E^\star$ obtains. ■

From the above proof, it is clear that knowledge that E may always serve as the self-evident event. To make this clear, I will express this result in the following lemma:

Lemma 4.4: Suppose (Ω, P) satisfies (P1) and (P2) *and* S knows E. Then the event $K(E)$ is self-evident for S.

Proof: See the proof for Theorem 4.1. ■

To understand the implications of these results, it is important to first reflect on what it means for knowledge to be self-evident. Suppose an agent S knows the empirical event E, "it is sunny outside". By Theorem 4.1 and Lemma 4.4, the event "$K(E)$" is then self-evident. Presumably, there are many ways in which $K(E)$ could obtain. S could know that it is sunny outside when she directly observes it, S could know that it is sunny outside when she hears it on the radio, S could know that it is sunny outside when her friend tells her it is sunny outside, etc. But because the event $K(E)$ is self-evident, Theorem 4.1 and Lemma 4.4 entail that in any of these ways S *could* come to know E, S does know E.

Above I suggested that empirical or contingent events should not be "nested" with respect to an agent's possibility function. What this means is that there are

ways in which an empirical or contingent event might occur without us knowing that it occurs. When an event is not nested in this way it cannot be self-evident. But the knowing of an event is, itself, an empirical or contingent event. At least it seems to be. Yet, if the KK principle is true (so (P2) and (K2) are true), then the knowing of an event does not have the familiar features of an empirical or contingent event. By Theorem 4.1 and Lemma 4.4, if S knows E (and the KK principle holds) then in any way in which $K(E)$ could obtain, S knows that it obtains.

The reason we should think that many standard empirical or contingent events like "it is sunny outside" are not self-evident is because we think there *can* be "evidential breaks" between the event happening and our knowing that it happened. The event "it is sunny outside" is not self-evident for me because one way for the event to obtain is when it is sunny outside and I'm camped out in the basement of the library reviewing old philosophy journals. In such an event there is a "break" between my evidence or information and the event. But, again, if the event $K(E)$ were self-evident, then in any way I *could* ever know E, I *would* know that I know E. If the event $K(E)$ were self-evident there would be, in principle, no possible way to break my evidential or informational access to knowing that I know. But because knowing a proposition appears, at least on the surface, to closely resemble an empirical or contingent event, it should seem implausible that knowledge is self-evident. Such considerations do not settle the issue, but I hope that they provide an initial reason to doubt the KK principle.

There is more to be said against the KK principle. Self-evident events provide a kind of epistemic "inverse black hole". In a black hole, nothing can escape and the entire region condenses to a point of infinite density. Self-evident events, instead, are wellsprings—from an epistemic point of "infinite density", knowledge flows forth, and knowledge of this knowledge, and knowledge of this further knowledge... *ad infinitum*. If an event E is self-evident for agent S, S knows E whenever E obtains. But the "epistemic density" of a self-evident event provides that S knows *this*—that

is, when E is self-evident, S knows that S knows E whenever E obtains. Such deep self-knowledge is implausible for real, non-ideal agents, as I'll argue next.[34]

The lemma below formally captures the idea that if knowledge is self-evident, then an agent possesses deep self-knowledge:

> **Lemma 4.5**: If $E^\star$ is self-evident for S, $\omega \in E^\star$ and (P1) and (P2) hold, then S knows that $E^\star$ is self-evident.
>
> **Proof**: Suppose $E^\star$ is self-evident. If $E^\star$ is self-evident then $E^\star \subseteq K(E^\star)$. By (K1) (which follows from (P1)), $K(E^\star) \subseteq E^\star$. So $K(E^\star) = E^\star$. But by (K2) (which follows from (P2)), $E^\star \subseteq K(K(E^\star))$. Because $\omega \in E^\star$, $\omega \in K(K(E^\star))$. (And $K(K(E^\star))$ is "S knows $E^\star$ is self-evident".) ■

What Lemma 4.5 establishes is that the KK principle (along with the factivity of knowledge) entails that agents *know* that their knowledge of E is self-evident, when they know E. But such a result is unacceptable for real, non-ideal agents, because we are never in an epistemic position to have such self-knowledge. Recall that in section 4.5 above, I argued that cognitive limitations such as perceptual discrimination limitations (Williamson's Mr. Magoo example), failures to record unremarkable events, and information processing limitations result in violations of the KK principle. Said differently, these cognitive limitations block higher-order and iterated knowledge. But what Theorem 4.1, Lemma 4.4, and Lemma 4.5 jointly entail is that if an agent knows any empirical event such as "it is sunny outside" and the KK principle holds, then she knows that these cognitive limitations *are not present*—that is, she knows that she *does not* suffer from limited powers of perceptual discrimination, that she *does not* suffer from memory failures for unremarkable events, and that she *does not* suffer from any processing errors, *or any other relevant cognitive limitation that*

[34] I will also argue in Chapter 5 that such deep self-knowledge is also implausible for ideally rational agents.

would impede higher-order knowledge. Such a list is open-ended—there are probably many cognitive limitations that I did not model in section 4.5 that could impact higher-order knowledge. The point is that because the KK principle requires deep self-knowledge, it entails that an agent knows of herself that there are no relevant cognitive limitations that block higher-order knowledge. We are not in a position to have such deep self-knowledge about the precise workings of our cognitive architecture, so we are not always in a position to know that we know (when we know).

There is another way to explain some aspect of the implausible epistemic requirements imposed by (P2) (and hence (K2) and the KK principle). In reviewing Lemma 4.1, it is clear to see that the first few steps of the proof establish the following:

> **Lemma 4.6**: If (K2) holds for agent S's knowledge operator K, then $P(\omega) \subseteq K(P(\omega))$.
>
> **Proof**: We know that $P(\omega) \subseteq P(\omega)$, trivially. By (K2) and the set-theoretic definition of knowledge it follows that $P(\omega) \subseteq K(P(\omega))$. ■

What Lemma 4.6 shows is that the KK principle entails a condition between epistemic possibility and knowledge: namely, that if some state is epistemically possible for S then S knows that the state is epistemically possible. Such a result is unacceptable for real, non-ideal agents. There are several ways to argue that this result is unacceptable. First, there are many possible states of which I am unaware. For all I knew five minutes ago, Charles Édouard Jeanneret (Le Corbusier) was born in Switzerland. Certainly, I did not know he was not, nor did I think he was not, nor did I have evidence that he was not born in Switzerland. In fact, five minutes ago I had never heard of the French architect Le Corbusier. Five minutes ago I did not know it was possible that Le Corbusier was born in Switzerland because I had no concept of Le Corbusier (I had never entertained any thought of Le Corbusier, heard

his name, etc.). But because knowledge is factive, the restriction (P1) requires that the actual state of affairs is epistemically possible for me. And because Le Corbusier was born in Switzerland, this is possible for me. So, such a case provides for the denial of the consequent of Lemma 4.6. So, by Lemma 4.6 and *modus tollens*, (K2) does not hold. Further, we are not in an epistemic position to know every possible state. One reason for this is that our unawareness of the possibilities stems from our lack of empirical access to information. We are necessarily empirically limited beings, so there will always be states of which we are unaware.[35] Hence, we are not in an epistemic position to satisfy (K2).

A second way to argue that the KK principle entails an unacceptable result is to note that real, non-ideal agents cannot enumerate every possible state. Real agents always have an incomplete picture of their state space. Beyond unawareness, there are possible states of which I cannot conceive. It is, of course, difficult to give examples of *these* states—if I could, then it seems that I could conceive of them. Better to consider historical figures. It seems reasonable to think that Thales of Miletus did not have the conceptual resources to conceive of principles of modern quantum electrodynamics, were they presented to him. So, it seems, Thales could not know such a principle were possible, even if it were explained to him by a modern day expert. Most likely, he would not understand the content of the utterances by the expert (even if such utterances were translated into his language). In this regard, Thales would be epistemically similar to a young child. Were I to explain a principle of quantum electrodynamics to a young child, the young child could not understand what is being asserted, and the young child would not know that the principle is possible. To the young child, it would be as if I uttered gibberish. Again, this is a phenomenon beyond unawareness—it concerns the limits of conceptualization. But,

[35]Could any human agent with enough time and mental resources learn the name of every human being that ever had a name? It seems highly doubtful, for the historical record is incomplete and impoverished. So there will always be people of which I am unaware.

such a principle of quantum electrodynamics *is* epistemically possible for the young child (and it was for Thales). Surely, the young child does not have evidence against the principle. Nor does the child know the principle is false. These cases provide for a denial of the consequent of Lemma 4.6. So, by Lemma 4.6 and *modus tollens*, (K2) does not hold. Further, as was the case above, we are not in an epistemic position to enumerate every epistemic possibility (due to inherent conceptual resource constraints), so we are not in an epistemic position to satisfy the KK principle.

Both these above remarks provide examples of how real, non-ideal agents do not have the deep self-knowledge that is required if the KK principle is true. Knowing that various knowledge claims are self-evident for one is to know deep features of one's cognitive architecture, that one does not make various *relevant* information processing errors. We do not have such knowledge. Always knowing what is epistemically possible requires deep awareness and conceptual resources that we do not possess. The KK principle makes cognitive demands beyond which we can meet, so the KK principle is false. Further, we are never in a position to satisfy these demands, so we are not even in a position to satisfy the KK principle.

Briefly, I want to address the implications of these arguments for ideally rational agents. Ideally rational agents will also always be unaware of aspects of their state space. As they are standardly understood, ideally rational agents are logically omniscient (they do not make mistakes in reasoning), but they are not empirically omniscient. The above arguments against the KK principle show how the KK principle actually entails powers of limited empirical omniscience. So, assuming that an agent satisfies KK is not merely to assume that she makes no mistakes in reasoning. I will revisit this claim in Chapter 5.

4.6.1 Remarks on the Argument

I've used the results from section 4.5 in conjunction with Theorem 4.1, Lemma 4.4, Lemma 4.5, and Lemma 4.6 to argue that the KK principle is false and also that

KK_{PTK} is false. Due to cognitive limitations, we do not always know that we know (when we know), nor are we always in a position to know that we know (when we know).

Lemma 4.6 showed one implausible implication of (P2) and (KK). I gave two arguments to think that just because some state ω is epistemically possible it need not be the case that we know it is possible. To show that (P2) (and, hence, (KK)) is really responsible for this consequence, it is worthwhile to consider a case where (P2) fails. Recall the epistemic possibility function from section 4.5.1, where agent S was unable to process and record "unremarkable events". S's possibility function was given as follows: $P(1) = \{1, 2\}$, $P(2) = \{1, 2, 3\}$, $P(3) = \{2, 3\}$. In this case the consequent of Lemma 4.6 does not hold; that is, $P(\omega) \not\subseteq K(P(\omega))$. To see this, consider the possibility set for S at 1: $P(1) = \{1, 2\}$. Now, at which states does S know $P(1)$? From the set-theoretic definition of knowledge, S only knows $P(1)$ at 1. So, $K(P(1)) = \{1\}$. But, clearly, $P(1) \not\subseteq K(P(1))$ because $2 \in P(1)$ but $2 \notin K(P(1))$. So, given S's possibility function, S does not know what is epistemically possible at 1.

The KK principle requires more than just higher-order knowledge: it requires deep self-knowledge. Human agents are never in an epistemic position to acquire this deep self-knowledge so human agents are not always in a position to know that they know, when they know.

4.7 Conclusion

I've suggested that Williamson's argument relies on inessential premises and that the fundamental issue concerns the structure of Mr. Magoo's possibility function and its implications for knowledge. Mr. Magoo's possibility function is such that Mr. Magoo need not always know that he knows, when he knows some proposition.

Further, I've shown how to extend the kind of argument Williamson suggests in order to accommodate other cognitive limitations. The formalism of information

structures provides a way of articulating and describing cognitive limitations and their implications for knowledge, higher-order knowledge, and epistemic possibility.

CHAPTER 5

INTERPRETING FORMAL MODELS

5.1 Introduction

In this chapter I discuss interpretive issues for the models of knowledge given in the previous chapters. I begin by considering a claim that is often made of idealized models of knowledge: ideally rationally agents partition their information space. I argue that ideally rational agents need not partition. I give a case which shows that agents may fail to partition because of perceptual processing limitations or lack of knowledge, and I argue that such shortcomings are not failures of rationality. Next, I show the relationship between partitional information structures and the existence of self-evident events. I show that if an agent partitions her information space then she knows an event E if and only if there is some self-evident $E^\star$ that obtains and $E^\star$ entails E. I discuss my previous arguments against the existence of self-evident events, and I argue that non-ideal agents do not partition their information space. I also explore the connection between the notion of awareness I discuss in Chapter 3 and epistemic possibility. I show how a notion of epistemic possibility that captures our "bounded" view of cognition has resources to provide an account of knowledge similar to the sentential awareness model of knowledge I gave in Chapter 3. As such, I show that there is a direct connection between epistemic possibility and failures of closure for knowledge. I conclude by discussing our inherent unawareness of what we know about what we know.

5.2 Rationality and Iterated Knowledge

In Chapter 4 I argued that the description of Mr. Magoo provides a case where an agent can know some proposition but fail to know that he knows. Modeling the epistemic features of Mr. Magoo in the set-theoretic model of knowledge shows *why* he fails to always know that knows—because Mr. Magoo's epistemic possibility function is not transitive (i.e., his possibility function does not satisfy (P2)), he fails to always know that he knows, when he knows. In this section I will discuss the broader relationship between rationality and iterated knowledge failures.

In many purposefully idealized models of knowledge, it is common to take the epistemic possibility relation to *partition* the set of states Ω. A partition $\mathcal{P}$ of a set S is a set $\{S_1, \ldots, S_r\}$ of nonempty subsets of S such that the members $S_j \in \mathcal{P}$ are pairwise disjoint and collectively exhaustive (i.e., $\bigcup S_j = S$).[1] One way to ensure that the epistemic possibility relation partitions the state space is to require that the possibility relation is reflexive, symmetric, and transitive. As such, it is often said that "rational agents partition the state space".[2] Why, then, do we have a link between rationality and partitional possibility functions? And what should we say about cases similar to Mr. Magoo and others that I previously raised in Chapter 4?

Before I discuss some cases in detail, I want to mention that the notion of rationality at play here essentially concerns epistemic possibility. To see why, consider a kind of flat-footed argument for denying the KK principle. Some have tried to argue that the KK principle is false because there are (or, at least, might be) cases of agents that know some proposition, but don't know that they know this because they lack

[1] Informally, a partition decomposes a set into a collection of non-overlapping "cells", such that any member of the original set S belongs to one and only one cell.

[2] For related discussion, see Binmore (1992), pp. 447–448. Geanakoplos (1992), p. 73, describes the content of common knowledge of rationality assumptions in the following way: "only the facts that the agents are rational—that is, their knowledge is given by partitions ..." (is common knowledge). See also Lipman (1995), p. 48.

the relevant conceptual resources for higher-level knowledge (thereby preventing iterated knowledge claims). I don't think these kinds of arguments are successful (or illuminating), but suppose they work to reject the KK principle. It seems that these arguments will not have the ability to explain the alleged connection between partitioning and rationality.[3] Often these arguments involve non-human animals, and the thought is that these animals can know various propositions ("the food is over there"), but they don't know that they know this. Because these animals violate the KK principle (K2), we know that their epistemic possibility relation is not transitive and, hence, they do not partition the state space. But should we say that such animals are irrational? At least, it seems that these animals do not make mistakes about their evidence or about what they know. They may have conceptual deficits, but this doesn't explain the intuition that it would be more rational for them to partition their state space.

Why, then, might partitioning be rational? Consider the case of Mr. Magoo again. Mr. Magoo considers a tree in the distance and wonders about its height. Mr. Magoo has decent vision, but the tree in question is somewhat far away. His vision is such that he cannot tell to the nearest inch how tall the tree is just by looking. So, supposing he has no other sources of information at hand, Mr. Magoo doesn't know how tall the tree is to the nearest inch. As Williamson (2000) describes him, "even if he so judges [the tree to be i inches tall] and in fact it is i inches tall, he is merely guessing; for all he knows it is really $i-1$ or $i+1$ inches tall…".[4] In Chapter 4 I suggested that the correct way to model Mr. Magoo's perceptual deficit was with the following possibility function: $P(i) = \{i-1, i, i+1\}$, where i denotes the state where the height of the tree is i inches. It is easy to show that Mr. Magoo's possibility function is not partitional—this is because his function P

[3]To be clear, in what follows I will argue that rational agents need not partition. But many have assumed otherwise. Those that endorse the flat-footed argument against the KK principle will not have the resources to explain this alleged connection.

[4]Williamson (2000), p. 115.

violates transitivity.

Now, necessarily, non-transitive possibility functions will have "overlap" when an antecedent condition is met; there will be some state s in more than one information set (i.e., there is some s such that $s \in S_i$ and $s \in S_j$ where $i \neq j$).[5] How do we interpret an agent's epistemic position for this "overlap" state s? By hypothesis, such a state is a member of at least two different information sets. A natural interpretation of such a situation is that the agent is failing to notice some piece of evidence. What might the agent be missing?

Consider the case of Mr. Magoo once again. His possibility function is such that $P(2) = \{1, 2, 3\}$ and $P(3) = \{2, 3, 4\}$. What information might Mr. Magoo have ignored? One suggestion is that Mr. Magoo doesn't know the structure of his own possibility function. Here's why: if Mr. Magoo knew the structure of his own possibility function, he would know that if it were possible that the tree was either 1, 2, or 3 inches tall, then the actual state would be 2, and he would update his possibility function to $P(2) = \{2\}$. But we've stipulated that this updated possibility function *does not* describe Mr. Magoo. So, it seems that he doesn't know the structure of his own possibility function. Yet, Mr. Magoo *does* know what is possible at 2. Because $\{1, 2, 3\} \subseteq \{1, 2, 3\}$, $\omega \in K(E)$ (where E is the event that the tree is either 1, 2, or 3 inches tall) at 2. And Mr. Magoo *would* know what is possible were any other state to obtain. So, we are forced to say that Mr. Magoo doesn't know the structure of his possibility function, even though Mr. Magoo knows what is possible at a state.

Does it seem irrational to not know one's possibility function? It doesn't. To know one's possibility function it is required that one know what *would* be possible were any other state to obtain. Yet this does not seem like a requirement of rationality. Surely, it would be ideal to know one's possibility function, but this doesn't

[5]Recall that a relation R is non-transitive when it violates the following condition: if Rab and Rbc then Rac.

have implications for epistemic rationality—it would be ideal (or, at least, *more ideal*) to know more than we currently know. Yet this does not bear on the issue of epistemic rationality. In a broad sense, epistemic rationality is about responding correctly to one's evidence, not about how much knowledge an agent possesses.

Consider another case. Suppose Ms. Magoo, who is also myopic, examines the readout on her electricity meter. Suppose that when the meter reads "1345" she cannot tell whether the meter reads "1345" or "1346", but when the meter reads "1346" she can discern exactly what it says. Ms. Magoo's possibility function is then represented by the following: $P(1345) = \{1345, 1346\}$, $P(1346) = \{1346\}$. Clearly, her possibility function is not partitional. Should we judge Ms. Magoo to be irrational? When she looks at the meter and it reads "1345", she cannot tell whether it reads "1345" or "1346". In this state, given her sensory evidence and everything she knows, the meter reading "1345" *seems just the same as* its reading "1346". Given her sensory evidence, these two states are indistinguishable. Now, one might be tempted to think that this is a rational failure on Ms. Magoo's part, for while "1345" seems just the same as "1346" when the state is 1345, "1345" does not seem just the same as "1346" when the state is 1346. The thought is that the relation "seeming just the same as" *ought* to be symmetric—generally, if "x" is indistinguishable from "y", then "y" ought to be indistinguishable from "x"—and Ms. Magoo violates this symmetry property. But what such an objection misses is that Ms. Magoo's evidence is *different* between the states 1345 and 1346. A plausible explanation for Ms. Magoo's situation is that when she observes a 5 she cannot tell whether it is a 5 or a 6, but when she observes a 6 she *can* tell that it is a 6. This might be a quirk of her perceptual system, but it does not seem to be a rational failure.[6] Hence, Ms. Magoo provides another case of a seemingly rational agent with

[6] As it turns out, humans have difficulty identifying human faces when the face is flipped upside down (inverted on the horizontal axis). As well, it does not seem to be irrational to not recognize an inverted face.

a non-partitional information space.

To conclude, it *seems* that perfectly ideal agents partition their information space. And insofar as an agent is perfectly ideal, she is also epistemically rational (that is, not being epistemically rational is not perfectly ideal). But, as I've shown, there are epistemically rational agents who do not partition their information space. These non-partitional information spaces may arise from perceptual processing errors or lack of knowledge (see especially Chapter 4), but these features are not (necessarily) failures of rationality. It may be more ideal to lack perceptual deficits and have more knowledge (hence explaining why it is *ideal* to partition), but this does not entail that it is irrational not to partition.

5.3 Self-Evident Events and Partitional Information Structures

In Chapter 2 I gave one argument against the existence of common knowledge by showing that an event E is common knowledge if and only if there exists some self-evident event $E^\star$ with the property that $E^\star$ entails E. Recall that an event $E^\star$ is self-evident if and only if for all $\omega \in E^\star$, $P(\omega) \subseteq E^\star$. Informally, a self-evident event is such that whenever it occurs, the agent knows that it occurs. I argued that for events of interest, there are no self-evident events that feature the appropriate entailment relation. Hence, no common knowledge. As I will argue below, a similar case can be made against any agent having a partitional information structure.

There are good reasons to deny the KK principle and the negative introspection axiom (K3) when modeling non-ideal human agents.[7] In Chapter 4 I argued against the KK principle by modeling several realistic features of human cognition. I also showed that reasons for denying the KK principle are reasons for denying the negative introspection axiom because the restrictions (P1) and (P3) on epistemic possibility relations entail (P2). Since denying (P2) is equivalent to denying either

[7] I first gave a presentation of the KK principle (K2) and the negative introspection axiom (K3) in Chapter 1.

(P1) or (P3), and because (P1) is true for any account of knowledge, we have reason to deny (P3) and, hence, deny negative introspection.

Reasons against the KK principle and the negative introspection axioms provide reasons against a partitional model of knowledge and partitional information structures for non-ideal agents. But a more general case can be made against partitional information structures. As it turns out, if the information structure (Ω, P) is partitional (i.e., induces a partition on Ω) then an agent knows the event E if and only if there is some self-evident $E^\star$ that obtains and $E^\star$ entails E.

> **Theorem 5.1**: Suppose (Ω, P) induces a partition on Ω. Then S knows that E if and only if there is some self-evident $E^\star$ that obtains and $E^\star$ entails E.
>
> **Proof**: Suppose S knows E. This means that $\omega \in K(E)$. Let $E^\star = K(E)$. Because (Ω, P) induces a partition, (P1) and (P2) hold, and so (K1) and (K2) hold. By (K1), $E^\star \subseteq E$, so $E^\star$ entails E. Next, I will show that $E^\star$ obtains and is a self-evident event. By (K2), $K(E) \subseteq K(K(E))$. But because $E^\star = K(E)$, this means that $E^\star = K(E^\star)$, so $E^\star$ is a self-evident event. Because $\omega \in K(E)$, it must be that $\omega \in E^\star$ (by $E^\star = K(E)$), so $E^\star$ obtains.
>
> For the other direction, suppose there is a self-evident event $E^\star$ such that $\omega \in E^\star$ and $E^\star \subseteq E$. By (K0), since $E^\star \subseteq E$, it must be that $K(E^\star) \subseteq K(E)$. And since $E^\star$ is self-evident, $E^\star \subseteq K(E^\star)$. Putting this together, we have that $\omega \in E^\star$ and that $E^\star \subseteq K(E^\star) \subseteq K(E)$, so it follows that $\omega \in K(E)$, so S knows that E. ■

In Chapter 2 and Chapter 4 I argued against there being any self-evident events of interest.[8] I will rehearse some of these arguments again. For an event to be self-

[8]The event Ω, the logical truism, is self-evident. But this is a trivial case. By "events of interest" I mean "informative" events, and not trivial events.

evident to agent S, S must know E whenever E occurs. Consider the event "the coffee cup is on the desk". Might this event be self-evident? Surely not. There are many ways the coffee cup could be on my desk without my knowing it. Simply, the coffee cup might be on my desk even though I don't notice that it is (suppose the cup is buried in philosophy books and paperwork). Instead, what about the event "I am appeared to redly"? It seems as though one may be appeared to redly even when one does not *notice* that one is. What about when one *believes* that one is appeared to redly? As Pollock and Cruz (1999) have argued, one may be mistaken about how one is appeared to, even when one has a corresponding appearance belief.[9]

It is worth noting how epistemically demanding self-evident events are. Foundationalists about epistemic justification have had a long history of searching for a class of what may be called "basic beliefs" to provide a secure substructure on which to justify all other beliefs. Though there are different ways to explicate the notion of a basic belief, the essential idea is that a basic belief is such that if it is held then it is true (i.e., it is impossible to hold the belief and be mistaken). Most epistemologists deny that there are any basic beliefs (or, they deny there are enough to serve the proper justificatory role). But notice that self-evident events reverse the order—with a basic belief, if it is believed then it is true. But a self-evident event is such that if it is true, then it is known. Basic beliefs attempt to provide certainty, but self-evident events seem to require a deep and ever-present awareness. Not only are we never wrong about the truth of a self-evident event (by the factivity of knowledge), we never miss a self-evident event when it happens. As I mentioned in Chapter 2, self-evident events are "luminous"—we are in a position to know that they obtain whenever they obtain. Williamson (2000) gives a forceful argument that there are no luminous events.[10] Because every self-evident event is a luminous event, Williamson's argument also counts against the existence of self-evident events.

[9]See Pollock and Cruz (1999), pp. 58–59.

[10]See especially Williamson (2000), Chapter 4.

By Theorem 5.1, supposing that (Ω, P) induces a partition on Ω, if we know any event E then there is some self-evident event $E^\star$ such that $E^\star$ entails E. Above, I gave reasons to doubt that there are any self-evident events. But there are equally strong reasons against thinking that self-evident events could serve in the required entailment relation. Consider many of the things we take ourselves to know. I know there is a coffee cup on my desk. Perhaps, I know that there is a coffee cup on my desk because it looks like there is a coffee cup on my desk. But, clearly, that it looks like there is a coffee cup on my desk does not *entail* that there is a coffee cup on my desk. It is consistent with its looking like there is a coffee cup on my desk that I am hallucinating a coffee cup on my desk. Moreover, many of the empirical propositions we routinely take ourselves to know are not entailed by other things we know. Mathematical and theoretical knowledge aside, entailment is rarely able to provide the epistemic support for what we know. These considerations, in light of Theorem 5.1, suggest that appropriate models of knowledge for non-ideal agents *should not* be partitional models.

However, there is more to be said. From the proof of Theorem 5.1, it is clear that the only necessary assumptions on P are that (P1) and (P2) hold.[11] These assumptions on P are, effectively, that P is reflexive and transitive. Because the reflexivity of P is equivalent to requiring that knowledge is factive (K1), we have good reason to endorse (P1). So, it seems, (P2) ought to be rejected if the consequences of Theorem 5.1 are to be avoided. This observation is consonant with my argument against the KK principle from Chapter 4—if knowledge is factive and satisfies the KK principle, then if an agent knows some event E (or proposition p), she knows a self-evident event that entails E. I argued in several places against the existence of self-evident events, so these reasons count in favor of denying the KK principle.

Now, one may observe that the proof of Theorem 5.1 requires the closure condition K0. I argued against several forms of closure in Chapter 3, so it is worthwhile

[11] Recall that (K0) is true in every information structure (Ω, P).

to consider the implications of denying closure for Theorem 5.1. First, closure was only used in the second direction of the proof. So, even without closure, it is true that if (Ω, P) induces a partition on Ω, then if S knows E, there is some self-evident event $E^\star$ that obtains and $E^\star$ entails E. This direction of the proof, alone, is enough to provide reason to reject partitional models of knowledge, in general, and (K2) in particular. There are no self-evident events of interest (i.e., there are no non-trivial self-evident events), so we have good reason to reject the KK principle.

5.4 Epistemic Possibility and Awareness

The core idea in the set-theoretic model of knowledge is that if S excludes all non-E states, then S knows E. Consider a concrete example. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$, representing the outcome of a roll of a six-sided die. Suppose that the actual state is 1 (the die roll resulted in a 1), and that Shane gathers evidence about the die roll such that his possibility function features the following: $P(1) = \{1, 2, 3\}$. His possibility function is supposed to model a case where it is epistemically possible for him that the die either landed on a 1, a 2, or a 3. One informal gloss on this condition is that, given what he knows, the die might have landed on 1, 2, or 3. Intuitively, Shane does not know that the die landed on 1. Why? For all Shane knows, the die might have landed on a 2, so he does not know that it landed on a 1. Let E denote the event "the die landed on 1" (so $E = \{1\}$). Because $P(1) \nsubseteq E$, the model agrees that Shane does not know the die landed on 1. It also seems intuitive that Shane knows that the die roll was less than 5. For all he knows, the die might have landed on a 1, 2, or 3, but given what he knows the die couldn't have landed on a 5 or a 6. The model also gets this simple case correct. Let F denote the event "the die roll is less than 5" (so $F = \{1, 2, 3, 4\}$). Because $P(1) \subseteq F$, Shane knows that the die roll is less than 5 when the true state is 1.

The set-theoretic model of knowledge makes clear the close connection between epistemic possibility and knowledge. But it leaves several important questions unan-

swered. What is epistemic possibility? Is it primitive (unanalyzable)? Can we be mistaken about what is epistemically possible? What is the connection between epistemic possibility and evidence?

To address these questions, I should first revisit a distinction I made in Chapter 1. The set-theoretic model of knowledge is a *model*, not an *analysis* of knowledge. Because the notion of a model is more involved, it is helpful to start with analyses of knowledge. For many philosophers, the goal of an analysis of knowledge is to provide necessary and sufficient conditions for knowledge that comport with a background of intuitive judgements about cases (e.g., ordinary cases, Gettier cases, barn facade cases, etc.). Ideally, the proposed individual necessary and sufficient conditions are "conceptually simpler" so that the analysis is informative or elucidatory. The aim of a *model* of knowledge, however, is more modest. For my purposes, models of knowledge provide tools meant to reveal conceptual connections and logical implications. Necessarily, models are idealizations (though some are comparably more or less idealized). As Ariel Rubinstein says of economic models, "the crowning point of making microeconomic models is the discovery of simple and striking connections between concepts (and assertions) that initially appear remote".[12] For example, I previously showed the relationship between the KK principle (knowing that you know) and a claim about the existence of self-evident events. Such a connection seems hidden without the use of a model of knowledge.

Most economic theorists who develop models of knowledge take the notion of epistemic possibility to be primitive to the model. That is, this notion is not analyzed in terms of simpler concepts, and it is a foundational concept for the model. In many cases, the notion is *not* interpreted as thoroughly normative, in that brute psychological features may influence what is counted as epistemically possible for an agent. In this way, models of knowledge are often presented as attempting to capture the agent's "viewpoint", or, in an imprecise sense, "what it is like (epis-

[12]See Rubinstein (1998), p. 191.

temically) to be the agent". For example, in Chapter 4 I gave a model of an agent who notices an event when it happens but does not notice that something does not happen when it does not happen. For a concrete example, consider an exchange between Sherlock Holmes and Colonel Ross:

> "Is there any other point to which you would wish to draw my attention"?
>
> "To the incident of the dog in the night-time".
>
> "The dog did nothing in the night-time".
>
> "That was the curious incident", remarked Sherlock Holmes.[13]

Suppose that Colonel Ross would notice the dog barking when it occurred, but that he would not notice that it didn't occur when it didn't occur. Let the states of the world be $\Omega = \{B, NB\}$, where "B" represents the state where the dog barks and "NB" represents the state where the dog does not bark. Colonel Ross's possibility function is then the following: $P(B) = \{B\}$, $P(NB) = \{B, NB\}$. That is, when the dog barks, Colonel Ross does not take it as possible that the dog didn't bark. But when the dog does not bark, Colonel Ross takes it as possible that the dog barked and as possible that the dog didn't bark. When the dog didn't bark, we model Colonel Ross *as if* he received no information as to whether the dog barked.

Now, there is a sense in which one might argue that Colonel Ross *ought not* take it as possible that the dog barked when the dog did not bark. One might suggest that Colonel Ross had the information or evidence that the dog did not bark, even though he took no notice of this information. Such a view seems to embrace an externalist view of information or evidence, and the model I am considering is, in this particular sense, an internalist model. From Colonel Ross' perspective, his not noticing that the dog did not bark is equivalent to receiving no information about the dog's barking.

[13]Doyle (1901).

Again, in the models of knowledge I am considering, the notion of epistemic possibility is not interpreted as thoroughly normative. What the agent takes as possible is not necessarily what the agent *ought* to take as epistemically possible.[14] Indeed, this is one of the model's virtues. Given Colonel Ross' failure to notice the dog's not barking, and his corresponding epistemic possibility function, it can be shown that this mistake *does not* prevent him from knowing that he knows (when he does). That is, it can easily be shown that Colonel Ross does not violate the property (P2), and this is enough to establish that he knows that he knows (when he knows). It can, however, be shown that Colonel Ross' possibility function does violate (P3), so there are cases when Colonel Ross doesn't know that he doesn't know (which is intuitively correct). The implication from the model, then, is that failing to take account of information does not, in principle, impact our knowledge of our knowledge.

To answer the above questions, I take epistemic possibility to be primitive to the model. The models of knowledge I consider aim to capture an agent's "view", however mistaken it may be, and an agent's corresponding knowledge. Can we be mistaken about what is epistemically possible? It depends. Again, there is a sense in which it is *not* epistemically possible for Colonel Ross that the dog barked when it did not. This sense may be explained in terms of an externalist conception of evidence or information. The thought is that even though Colonel Ross did not take account of the information that the dog did not bark, this information was in principle available. To contrast, I am interested in modeling how agents view the world, perhaps "from the inside", and, hence, there is also a strong sense in which it is epistemically possible for Colonel Ross that the dog barked (when it did not). Depending on one's view of evidence, an agent may have evidence for an event but not take the event to be epistemically possible.

[14]Likewise, one may say that the states an agent excludes as possible are not necessarily those that she ought to exclude as possible.

5.4.1 Awareness

Once capturing an agent's "view" is taken to be a goal, several natural features of epistemic possibility become apparent. It is clear that our view of the world is necessarily bounded, in that we cannot conceive of or imagine many states of the world. For instance, take any distant celestial body such as a remote star, one that we could not identify with the naked eye. Call this object "X". Supposing that X exists, most humans are unaware of X. What this might mean is that most humans do not have a name for X, most humans have never had a thought about X, and most humans do not have a concept of X. At the other end of the cosmic spectrum, most humans are unaware of neutrinos. Again, most humans do not have a name for neutrinos (they have never heard of neutrinos), or have thought about neutrinos, or have a concept of neutrinos.

Our unawareness of the world provides a kind of bound on our view of the world. Surely, when we were younger there were things that we were unaware of but, through experience, have since become aware. Our awareness typically expands as we gain new evidence and information about the world. But there are always parts of the world of which we are unaware. These items may include distant astronomical objects (e.g., the satellite Triton), remote islands in the Philippines (e.g., the Babuyan Islands), obscure historical figures (e.g., Horace Walpole (1717–97), an English politician and author), specialists' tools (e.g., a tennis stringer's awl), among many others. The crossword and Jeopardy are good ways to find things of which we were previously unaware.

Given that I am interested in capturing an agent's view, it seems natural to impose the following restriction on epistemic possibility:

(A) If S is unaware of p, then p is not epistemically possible for S.

Though there are details to work out, I take it that (A) is intuitively plausible. The idea is that if an agent cannot conceive of some proposition then it is not possible

(for the agent) that the proposition is true. In partial support of (A), consider the case of King Henry I (1068–1135). It should be uncontroversial that Henry I was unaware of neutrinos. Likewise, it seems reasonable to assert the following: it was not epistemically possible for Henry I that neutrinos have half-integral spin. Or, perhaps more naturally: it was not possible for Henry I that neutrinos have half-integral spin. The rough idea for now is (I will expand on this point below), because Henry *could not* conceive of neutrinos, it wasn't possible for him that neutrinos have half-integral spin. If we take epistemic possibility to comprise an important but unresolved portion of an agent's view, unawareness provides a boundary to epistemic possibility.

Continuing with the above line of thought, there may be an intermediate necessary condition on epistemic possibility that can be invoked in support A. Consider the following:

> (B) If p is epistemically possible for S, then S can understand the content of p.

Why think that (B) is true? Perhaps, what is epistemically possible for S has to do with what *might* be the case, from the agent's perspective. But, it seems, if S cannot understand p, then it is not true that p might be the case, from the agent's perspective. When an agent cannot understand the content of p, this proposition does not express (for the agent) a way the world might be.

Now, if (B) is true, then we have reason to endorse (A). For, if S is unaware of p then S cannot understand the content of p. The case of Henry I provides an example. Because Henry I was unaware of neutrinos, he lacked the conceptual resources to understand the proposition "neutrinos have half-integral spin". Of course, in some instances, the very act of hearing some proposition may make one aware of (in some sense) the constituents of the proposition. Hence, there may be reason to include temporal subscripts to both (A) and (B). For instance, suppose that I had not heard

of the planet Rigel 7 at t (that is, at t, I was unaware of Rigel 7). If you ask me if I know that Rigel 7 is the seventh planet in the Rigel star system at $t + 1$, the very mention of Rigel 7 may be enough to make me aware of Rigel 7. It may be true, then, that at t it was not epistemically possible for me that "Rigel 7 is the seventh planet in the Rigel star system", but at $t + 1$ it was possible for me that "Rigel 7 is the seventh planet in the Rigel star system". It is unclear that the mention of neutrinos would be enough for Henry I to understand the proposition "neutrinos have half-integral spin". Surely, Henry I is also unaware of the property of having half-integral spin, and subatomic particles, and atomic particles, and particles.

It is important to note how (A) conflicts with other accounts of epistemic possibility. Consider what is often taken to be a beginning account of epistemic possibility:

> (C) p is epistemically possible for S if and only if p is consistent with everything S knows.[15]

According to (C), many propositions of which we are unaware are epistemically possible. For instance, suppose that in 4000 years scientists will discover that all matter is composed of what are called "particuli", and all particuli have the property of "negative alpha-delta charge". Supposing that we currently do not have evidence to the contrary, by (C), it is then epistemically possible for us that "particuli have negative alpha-delta charge". Insofar as there is a connection between our view of the world and epistemic possibility (a connection I mean to emphasize, at least), (C) is not restrictive enough. By (C), statements about arcane and obscure mathematical theorems are epistemically possible for any human because they are necessary. As well, many statements about the future are plausibly epistemically possible for us, by (C). Consider the proposition "Martha Bayberry is the sixty-fourth president of the United States of America". It seems that this propositions is consistent with

[15]See Huemer (2007), especially p. 124, for several common accounts of epistemic possibility. Two propositions are consistent if and only if they do not jointly entail a contradiction.

everything I know (at least, I don't know that she won't be president). But it is unclear what meaning "Martha Bayberry is the sixty-fourth president of the United States of America" has for me. Supposing there is some content to the proposition, it seems odd that this proposition was epistemically possible for me before I first wrote the above sentence. There is a strong sense in which whether Martha Bayberry's being president was not part of my view of the world (again, before I first wrote the above sentence).

To reiterate, I am interested in accounts of epistemic possibility that capture an agent's view of the world. As such, I think (A) captures something important about resource bounded agents' epistemic situation. To be clear, I am not giving a semantic account of epistemic modals such as 'might' or 'can'. Surely, there is much more to what we mean with epistemic modals than is represented by (A). Nonetheless, epistemic modals may require a similar constraint to (A) for a complete theory of their truth conditions. If I truly say "it might be raining in Tobago", it seems to either entail or presuppose that I am aware of Tobago. That is, in most cases I cannot utter the sentence 'it might be raining in Tobago' if I am unaware of Tobago.[16] Typically, when S is unaware of some object O, S does not have a word for O in her lexicon. And one cannot utter a sentence s if one does not have the words to express s in one's lexicon. Utterance requires articulation and articulation requires a lexicon.

Briefly, consider our epistemic modal attributions to others. It is rare to say, of someone else, that "it is possible for her that p". A more natural way to express a similar thought is with the construction "for all she knows". For instance, we may say "for all she knows, the birthday party will be on Thursday" or "for all she knows, the birthday party might be on Thursday". Yet, plausibly, it would be odd to make such a statement if the subject of the attribution lacked awareness of

[16]I can utter the sentence 'it might be raining in Bananaland' even though I intend for 'Bananaland' to denote a fictional place, but it is unclear whether this utterance has any content.

the constituent parts of the statement. Suppose I aim to make a statement about Charles, a typical four year-old boy. It seems that only in jest could I say, "for all Charles knows, the Higgs boson might be found at 1.4 teraelectronvolts (TeV)". Plausibly, I might say this to inform the listener that Charles knows nothing about particle physics, or that Charles has never heard of the Higgs boson particle. But it seems wrong to infer from my statement that Charles' evidence does not rule out finding the Higgs boson at 1.4 TeV.

5.4.2 Awareness and Awareness Structures

In this section I will show how a notion of epistemic possibility that is consistent with principle (A) above can provide the resources for a view about knowledge similar to the sentential awareness model of knowledge that I gave in Chapter 3. If (A) is true, then, in a broad sense, epistemic possibility is restricted by our awareness. By (A), if S is unaware of p, then p is not epistemically possible for S. It is easy to see how such a view about epistemic possibility might have implications for epistemic closure, similar to the sentential awareness model of knowledge. Suppose S knows p and that p entails q, but S is unaware of q. If we keep the set-theoretic definition of knowledge from Chapter 1, it is then easy to show that S does not know q (even though what he knows entails q). That is, because q is not epistemically possible for S, it is not the case that every epistemic possibility for S is a q state.

The above explanation is not perfectly correct. The set-theoretic model of knowledge addresses knowledge of events, but my remarks so far have involved knowledge and awareness of propositions. Accordingly, I need to explicate the notion of unawareness of an event. In Chapter 1 I showed how it is possible to translate between the set-theoretic model of knowledge (which does not contain an object language) and more traditional epistemic logics (which contain an object language). The idea was that a state is a "full description of the world" in some unspecified sense. Identify a state $\omega \in \Omega$ with a set of true propositions. Under this identification, the

notion of a full description of the world corresponds to, roughly, the propositional content of the collection of true propositions at that state.

I previously suggested that agents are aware and unaware of propositions. The natural extension of this idea to events is the following:

> *Event Unawareness*: S is unaware of the event E if and only if S is unaware of some proposition $p \in E$.

Now, Event Unawareness is not yet enough to produce failures of closure, similar to the sentential awareness models of knowledge. To get these results, endow an agent with her "awareness set" at each state ω:

> **Awareness Set**: the proposition p is a member of S's awareness set $\mathcal{A}(\omega)$ at ω (i.e., $p \in \mathcal{A}(\omega)$) if and only if S is aware of p at ω.

Next, modify the set-theoretic definition of knowledge to include a clause concerning an agent's awareness set.[17]

> **Set-Theoretic Definition of Knowledge with Awareness**: Let (Ω, P) be an information structure. The event E is known at state ω if and only if $P(\omega) \subseteq E$ and for every $p \in E$, $p \in \mathcal{A}(\omega)$.

The idea in the set-theoretic definition of knowledge with awareness is that a proposition is known (by S at ω) if two conditions are met: (i) every state possible for S is an E state and (ii) S is aware of E.

From the set-theoretic definition of knowledge with awareness it is easy to see why closure fails. Suppose that S knows E, and S knows that the event E entails F, but S is unaware of F. In this case, S doesn't know F because she does not satisfy the second condition of the set-theoretic definition of knowledge with awareness, but she does know E.

[17] This addition is similar to the condition (KN) in Chapter 3.

There is another way to get the properties of the set-theoretic definition of knowledge with awareness *without* invoking an awareness set or modifying the original set-theoretic definition of knowledge. Because we can identify a state ω with a set of propositions, there is already a "sentential" or language-like aspect to a state. To see how this is accomplished, however, it is important to be clear about the notion of a state.

Let any collection of *sentences* be a state. Two states ω and ω' are equivalent if and only if $\omega = \omega'$, where the relation '=' is interpreted as the identity relation over sets. Because states are now understood "sententially" (rather than in terms of propositions), intuitively equivalent states are no longer thought to be identical. For example, let 'p' be the sentence 'the die rolled is a 2' and let 'q' be the sentence 'the die rolled is (strictly) greater than 1 and (strictly) less than 3'. While it is reasonable to think that p will be true if and only if q is true, p and q are different sentences. Hence, if $\{p\} = \omega$ and $\{q\} = \omega'$, $\{p\} \neq \{q\}$.

Consider an example. A third party rolls a six-sided die and the agent S observes the roll. The die lands on 2. Let $\omega = \{p\}$, where 'p' is the sentence 'the die lands on 2', and let $\omega' = \{q\}$, where 'q' is the sentence 'the die roll is even'. Let S's possibility function be described as follows: $P(p) = \{p\}$. By the set-theoretic definition of knowledge, S knows ω (i.e., S knows the die lands on 2). By a traditional understanding of logical entailment, ω entails ω'—if the die lands on 2, then the die roll is even. But, by the set-theoretic definition of knowledge, S does not know ω'. To see why, it is straightforward to check that $P(p) \not\subseteq \omega'$. This is because $p \in P(p)$, but $p \notin \omega'$.

The above example is highly artificial. But the example shows how, by modifying the notion of state to a "sentential" notion, a strong form of closure fails. It may be that when we know that the die roll is a 2 we thereby know that the die roll is even. Perhaps we can understand this claim as one about sentential possibility: any time we consider the sentence 'the die roll is a 2' as possible we also consider the

sentence 'the die roll is even' as possible. But there are clearly cases where such a relation does not obtain. Probably many of us forget that 2 is the smallest prime. Suppose that this means we are unaware that 2 is the smallest prime.[18] In this case, it is not true that any time we consider the sentence 'the die roll is a 2' as possible we also consider the sentence 'the die roll is the smallest prime'.

The motivation for moving to understand a state as a collection of sentences is similar to the motivation for understanding our view of the world as bounded. Because there are many things of which I am unaware, there are many sentences of which I am unaware. There was a time in which I was unaware of Rigel 7. As such, I was unaware of whether Rigel 7 is the seventh planet in the Rigel star system. Though this may have been consistent with what I knew, such a possibility was not under my consideration in any way—it was not part of my "view" of the world.

In Chapter 3 I discussed the sentential awareness model of knowledge. For closure to fail, I gave condition (KN) on knowledge which essentially adds a further condition on knowledge: to know a proposition, one must know it in the sense described by the knowledge operator K *and* one must be aware of the proposition. But if we understand a state as a sentential construction, such a conjunction is unnecessary. Under a sentential notion of a state, agents have sentential knowledge, and such knowledge is not necessarily closed. And just as I considered various restrictions on $\mathcal{A}(s)$ in Chapter 3, there is room to add further restrictions on the notion of a state. In considering various agents, it may be appropriate to require that if the sentence '$p \wedge q$' is a member of the state ω then so is the sentence '$q \wedge p$'. Such requirements will have various implications for the model, revealing further connections between our view of agents as cognitively limited and what they know.

[18]The case could be modified with an assumption that we have never heard of prime numbers.

5.5 Knowledge of Awareness

I have previously argued that, as real, non-ideal agents, our view of the world is bounded, insofar as there are always objects of which we are unaware. At the same time, we often know this about ourselves. For instance, I know there is *something* of which I am currently unaware, even if I cannot say precisely what it is of which I am unaware. I know there are cities in the world that I have never heard of, I know there are historical figures that I have never heard of, and I know there are distant planets and solar systems of which I have never heard. Using technical jargon, I know my state-space is incomplete, yet I do not know the degree to which it is incomplete. In this section I will argue that this is also our relationship with respect to what we know—we are always unaware to some degree of what we know.

What do we know about what we know? I've already addressed two important aspects to this question when I previously discussed the knowledge axioms (K2) and (K3), given again below:

(K2): $K(E) \subseteq K(K(E))$,

(K3): $\neg K(E) \subseteq K(\neg K(E))$.[19]

In Chapter 4 I argued against (K2). And because the factivity of knowledge in conjunction with (K3) entails (K2), my arguments against (K2) can serve as arguments against (K3).[20] However, it is worth considering what we *would* know about what we know, were (K2) and (K3) true. If (K2) and (K3) were true, we would know *a lot* about what we know. Suppose that S knows E, the event that it is sunny outside. By (K2), S would know that she knows E, and S would know that she knows that she knows E, ... *ad infinitum*. Ignoring finitary models of knowledge (though see Chapter 4), (K2) provides for an infinite number of knowledge claims.

[19] I first discussed these axioms in Chapter 1. See also Chapter 4.

[20] I take the factivity of knowledge to be a requirement of any model of knowledge. I showed that the factivity of knowledge in conjunction with (K3) entails (K2) in Chapter 1.

Now suppose, instead, that S doesn't know that it is sunny outside. By (K3), S then knows that she doesn't know this. And by (K2), S knows that she knows that she doesn't know this, and S knows that she knows that she knows that she doesn't know this, *ad infinitum*. In an informal sense, if (K2) and (K3) were true, then we would know very clearly what we know and what we don't know. There would be no uncertainty about what we know and what we don't know.

But such a situation does not seem to be true for human knowers as real, bounded agents. We are often unsure of aspects of what we know and what we don't know. As an example, consider the previous case of Colonel Ross, given in section 5.4 above. From the example, Colonel Ross was charged to be on the lookout for possible suspects, yet he did not notice the "curious incident of the dog in the night". But he also did not notice that he did not notice the incident of the dog: as he suggested (when prompted), "the dog did nothing in the night-time". But, as Holmes pointed out, "that was the curious incident".

To model such a scenario, let $\Omega = \{a, b, c\}$.[21] In state a, the dog barks and Colonel Ross realizes there is a human intruder. In state b, the cat howls and Colonel Ross realizes there is a canine intruder. In state c the dog does not bark and the cat does not howl, and Colonel Ross does not notice that nothing has happened. Given the description of Colonel Ross, let his possibility function have the following structure: $P(a) = \{a\}$, $P(b) = \{b\}$, $P(c) = \{a, b, c\}$.

The intended interpretation of the story of Colonel Ross is that, at c, he is epistemically impoverished in some way: at c he seems to not know that he is at c, because he didn't notice that he didn't notice what was happening. From the set-theoretic model of knowledge, it is clear that at c Colonel Ross does not know that the dog barked, or that state a obtained (because $P(c) \not\subseteq \{a\}$). This is good because, at c, the dog did not bark, so Colonel Ross cannot know this. However, Colonel Ross does not know that he does not know *this*. To see why, consider the

[21] My presentation and modeling of the Colonel Ross example follows Dekel et al. (1998).

two events "a is not known" and "it is not known that a is not known":

$$\neg K(\{a\}) = \{b, c\}$$
$$\neg K \neg K(\{a\}) = \{a, c\}.$$

From above, taking the intersection of $\{b, c\}$ and $\{a, c\}$, we see that at state c, Colonel Ross does not know a *and* Colonel Ross does not know that he does not know a. In fact, given his possibility function, at c Colonel Ross has no positive knowledge of a at all. This property can be expressed by the following:

$$\{c\} = \bigcap_{i=1}^{\infty} (\neg K)^i(\{a\}),$$

where $(\neg K)^i(\{a\})$ means that the "not known" operator $\neg K$ is iterated i times. So, at c, Colonel Ross does not know a, Colonel Ross does not know that he does not know a, Colonel Ross does not know that he does not know that he does not know a, ... *ad infinitum*.

The above infinite iteration of not knowing a expresses Colonel Ross' impoverished epistemic situation at c. Because he took no notice of his situation, because he didn't notice that the dog didn't bark in the night, he was *unaware* of important information. Surely this is a plausible response to the Colonel Ross case. Indeed, such a response fits well with the approximate colloquial definition of "to be unaware" as "having no knowledge of a situation or a fact". Colonel Ross didn't know, didn't know that he didn't know, and so on: hence he was unaware.

Though this is a slightly different notion of unawareness from what I discussed in Chapter 3, I want to investigate this concept, unawareness of an event as "having no knowledge" of the event.[22] Formally, the idea is that an agent S is unaware of event E if and only if S doesn't know E, S doesn't know that S doesn't know E, ..., *ad infinitum*. Define an "unawareness operator" as an infinite conjunction of iterations of not knowing:

[22]In Chapter 3 I discussed what may be called "object awareness", which was closely related to the notion of having a name for an object in one's vocabulary.

Event Unawareness: $U(E) = \bigcap_{i=1}^{\infty} (\neg K)^i(E)$.

Again, the idea is that an agent S is unaware of some event E when S has no positive knowledge of E.

When Colonel Ross was on lookout, he didn't hear a dog bark, so he didn't know that a dog barked. But Colonel Ross also failed to notice that nothing happened with the dog, so he didn't know that he failed to notice. Regarding the dog's barking, Colonel Ross was unaware. Holmes, in contrast, knew that Colonel Ross didn't know that the dog barked. Holmes noticed that Colonel Ross didn't notice. And so Holmes was aware of Colonel Ross' unawareness.

Human agents seem more like Colonel Ross in this respect. We often fail to notice some event, and we fail to notice that we didn't notice. After all, it is usually cognitively difficult to notice that one did not notice something, when one didn't notice in the first place. Consider the celebrated example of inattentional blindness given by Daniel Simons and Christopher Chabris (1999). Subjects were instructed to watch a video of students dribbling a basketball and count the number of times that the ball was passed between students. However, while the students passed the basketball back and forth, a person in a gorilla suit walked across the video frame. Remarkably, many subjects never noticed the gorilla. Now, these students didn't know a gorilla walked through the basketball passing exercise. But it also seems plausible to think that these subjects didn't know that they didn't know this. That is, it is plausible to think that these subjects didn't know that they failed to notice the gorilla. And it is plausible to think that the subjects didn't know *this*. Indeed, the subjects are commonly described as being unaware of the gorilla.

5.5.1 Unawareness of Knowledge

In what is to follow I will argue that we are always unaware to some degree of what we know. Just as Colonel Ross was unaware of the "curious incident in the

night", real, non-ideal agents are always such that they do not know some iterative knowledge claim about what they know, and they do not know *that*, nor *that*, ... *ad infinitum*.

Begin with the reasonable assumption that for any event E (or proposition p), agent S either knows E or agent S does not know E. Further, assume that claims about what an agent knows or does not know have propositional content, so these claims can be known or not known. These two assumptions entail the existence of an infinite number of iterative knowledge claims for any agent S and event E. That is, for any event E, S either knows E or doesn't know E; and, if E is known, S either knows that S knows E or S does not know that S knows E, else either S knows that S doesn't know E or S does not know that S does not know E, ... *ad infinitum*.

One way to formalize the above idea is by way of a sequence. For any event E, there will be some sequence of knowledge claims that is true for S. One example might be the following:

(i) $K(E)$

(ii) $\neg K(K(E))$

(iii) $K(\neg K(K(E)))$

(iv) $K(K(\neg K(K(E))))$

(v) $\neg K(K(K(\neg K(K(E)))))$

(vi) $K(\neg K(K(K(\neg K(K(E))))))$

(vii) $\neg K(K(\neg K(K(K(\neg K(K(E)))))))$

$\vdots$

In this example, S knows E, but S does not know that S knows E. Further, S knows *this*: i.e., S knows that S does not know that S knows E. And so on.

What is at issue is the progression of the outermost knowledge operator at each iteration, the first knowledge operator at each new successive knowledge claim. In the above example, S knows E so there is a knowledge claim at (i). But S does not know (i), so there is a negated knowledge claim at (ii). Letting '0' denote a negated knowledge operator and letting '1' denote a non-negated knowledge operator, the above example can be represented as the following:

(Seq. 1) $\ldots 0101101E$.

Next, it is helpful to have a way to refer to points or positions on what I will call the "iterative knowledge sequence". Reading from right to left immediately to the left of the event E on (Seq. 1), call the number just to the left of the event E the "first point", call the number just to the left of the first point the "second point", and so on. Refer to the number immediately to the left of the $n-1^{th}$ point as the "n^{th}" point. From the above example, there is a 1 in the first point, a 0 in the second point, a 1 in the third point, and a 1 in the fourth point. Lastly, it is important to note that this sequence is infinite. That is, for any $n \in \mathbb{N}$, the n^{th} point will either be a 0 or a 1. This is because an agent will either know or not know any knowledge claim with $n-1$ iterated negated or non-negated knowledge claims about E.

Because (Seq. 1) represents an infinite number of knowledge claims, (Seq. 1) is true of S if and only if the following is true of S: S knows E, S does not know that S knows E, S knows that S does not know that S knows E, $\ldots$ *ad infinitum*. That is, (Seq. 1) is true of S if and only if the original sequence of knowledge claims given in the example is true of S.

As I previously stated, I will argue that we are always unaware to some degree of what we know. Using the sequence notation, this claim amounts to the following:

> **Unawareness of Knowledge**: for any event E, there exists some point n (the n^{th} point) such that every point $m > n$ is a 0.

What I am claiming is that for any event E and any non-ideal agent S, there will be some finite sequence of negated and/or non-negated knowledge claims about E that S does not know, S does not know that S does not know, S does not know that S does not know that S does not know, ... *ad infinitum*. To see an example, suppose that for S the following first four points of the iterative knowledge sequence are explicitly given below:

(Seq. 2) ...1011E.

If Unawareness of Knowledge is true, then there will be some n such that every point $m > n$ is a 0, so there will be some greatest n^{th} point that is a 1 such that

(Seq. 3) $\ldots 000000\underbrace{1\ldots1011}_{\text{n points}}E$.[23]

It is important to note that the claim

"$1\ldots1011E$"

as a component from (Seq. 3), is itself a knowledge claim about what S does and does not know about E. Now, from the example, it is this claim "$1\ldots1011E$" that S does not know, S does not know that S does not know, ... *ad infinitum*. If Unawareness of Knowledge is true, then in this example S is unaware of this higher-order knowledge claim ("$1\ldots1011E$") about her knowledge of E. Further, if Unawareness of Knowledge is true, then there will always be some knowledge claim for any agent S about the event E such that S is unaware of E.

Consider the following concrete example. Suppose that Andy knows that it is sunny outside. Further, suppose that she knows that she knows this and is always

[23]Of course, S may be unaware of E, such that S has no positive knowledge of E. In this case, while there would be no greatest point n that is a 1, Unawareness of Knowledge will still be satisfied. However, because (Seq. 2) contains a '1', if Unawareness of Knowledge is true then there will be some greatest n^{th} point that is a 1.

careful to consider what she knows and what she does not know. It might be the case that Andy has many iterations of knowledge about her knowledge of its being sunny (she knows that she knows that she knows...). But if Unawareness of Knowledge is true, such knowledge must eventually run out. If Unawareness of Knowledge is true, at some point in the consideration of ever higher orders of knowledge, Andy will be unaware of her iterated knowledge of its being sunny—that is, her positive knowledge of what she knows she knows will run out. And the same is true for Andy's knowledge of what she does not know. Andy might know that she does not know that it is *not* sunny outside, and she also might know that she knows *this*. But, again, if Unawareness of Knowledge is true then at some point in her consideration of higher orders of her knowledge, Andy will be unaware of her iterated lack of knowledge that it is not sunny.

Before I argue for Unawareness of Knowledge, I want to show what impact principles (K2) and (K3) have on the iterative knowledge sequence. Suppose that (K2) (the KK principle) is true. Then if there exists some point n on the iterative knowledge sequence that is a 1, then any further point $m > n$ will be a 1. Suppose, instead, that (K3) is true. Then if there exists some point n on the iterative knowledge sequence that is a 0, then any further point $m > n$ will be a 1. So, (K2) and (K3) jointly entail that there will always be some point n such that every point $m > n$ is a 1. Hence, (K2) and (K3) are jointly inconsistent with Unawareness of Knowledge.[24] So it follows that if Unawareness of Knowledge is true, then (K2) and (K3) cannot both be true.

Why think that Unawareness of Knowledge is true for all real, non-ideal agents? The first reason is that non-ideal agents do not have the cognitive resources to entertain every point of the infinite iterative knowledge sequence. So it seems that there should be some point on the infinite iterative knowledge sequence in which the

[24]Note, however, that the denial of (K2) and (K3) is not equivalent to a denial of Unawareness of Knowledge.

agent cannot parse and understand the expressed proposition and hence not know it, or any further concatenation of it. Consider the following knowledge claim:

> (1) S knows that S knows that S does not know that S knows that S does not know that S does not know that S does know that S knows that S does not know that S knows that S knows that S knows that S knows that S does not know that S does not know that S knows that S knows that S knows that E.

I don't fully understand the proposition expressed by (1). At some point in reading the sentence I lose track of what is claimed to be known. And because I don't understand the content expressed by (1), it seems beyond my cognitive ken to know that it is true.[25] Any outsider should judge that I do not know (1).[26] And likewise for a version of (1) about what I know. Even with time for careful study and proper concentration, it seems that I could never know iterative claims about what I know fifteen or sixteen times the length of (1). Perhaps it is at this point where I don't know the claim, I don't know that I don't know the claim, ... *ad infinitum*, in accordance with Unawareness of Knowledge.

However, matters are slightly more complicated. From my previous argument above, I actually know that I don't know (1). Yet because I cannot articulate the content of (1), I don't know exactly what it is that I don't know. This is a further example of my unawareness of what I know and what I don't know. I know that I don't know whatever is expressed by (1), but I do not know precisely what this is. Similarly, suppose that I successfully argue for Unawareness of Knowledge. Then I

[25]See the following paragraph for a further discussion of this claim. Of course, I could learn by testimony that (1) is true and thereby come to know (1). In this case there is a sense in which (1) is not beyond my ken. Yet one could easily substitute (1) for the event E to provide a plausibly unknown modified version of (1).

[26]See also the similar example I gave in Chapter 2 regarding the infinite common knowledge sequence.

know that there is some point in every infinite iterative knowledge sequence that is not known, but I do not know precisely where this boundary lies. Were I to know the location of this boundary, it would disappear. I know that I am unaware, but such knowledge is itself indefinite.[27]

Moreover, the claim that real, non-ideal agents cannot entertain, consider, parse, or understand knowledge claims such as (1) (and thereby cannot *know* such claims) need not be viewed as favoring either internalism or externalism about knowledge. I take it that this is fundamentally a claim about the limits of cognition. One may spell out these limits in terms of "information" or "reasons" or "evidence", but none of these interpretations is privileged given what I've said so far.

A second consideration for Unawareness of Knowledge concerns the sequence of values for the infinite iterative knowledge sequence. I will argue that real, non-ideal agents are never in a position to have infinite, positive higher-order knowledge for any event E. Suppose, instead, that for some event E, agent S is such that there exists some point n such that every point $m > n$ is 1. As an example, S might be such that the following is true:

$$\text{(Seq. 4)} \ldots 111\underbrace{0\ldots1011}_{\text{n points}}E.$$

By the lemma given below, I will argue that there cannot be non-ideal agents that possess such positive higher-order knowledge.

> **Lemma 5.1**: Let P be an information structure for S satisfying (P1), and let K be the knowledge operator induced from P. Suppose that S knows E, and that S knows that S knows E, ... *ad infinitum*. Then there exists some self-evident event $E^\star$ such that $E^\star$ entails E.

[27] Recall that in Chapter 3 I suggested an interpretation where all knowledge is indefinite. Because our knowledge of our state-space is incomplete, what we know when we know an event E is that we are in some "E-state".

Proof: Take $E^\star$ to be the intersection of all sets of the type $K(E)$, $K(K(E))$, $K(K(K(E)))$, ... Because it is true that $K(E)$, $K(K(E))$, $K(K(K(E)))$, ..., this means that $\omega \in K(E)$, $\omega \in K(K(E))$, $\omega \in K(K(K(E)))$, ..., so $\omega \in E^\star$. By (K1), which follows from (P1), $E^\star \subseteq E$, so $E^\star$ entails E. To show that $E^\star$ is a self-evident event, one must show that for any $\omega' \in E^\star$, $P(\omega') \subseteq E^\star$. Suppose $\omega' \in E^\star$. But because $E^\star$ is the non-empty intersection of all sets of the type $K(E)$, $K(K(E))$, $K(K(K(E)))$, ..., it is true that $P(\omega') \subseteq K(E)$, $P(\omega') \subseteq K(K(E))$, ..., so it follows that $P(\omega') \subseteq E^\star$. ■

I have argued in several places against the existence of non-trivial self-evident events (see section 5.3 and Chapter 2 and Chapter 4). I have also argued that there are no self-evident events that *entail* ordinary empirical events such as "it is sunny outside". By contrapositive reasoning in conjunction with Lemma 5.1, non-ideal agents never possess such positive higher-order knowledge. That is, non-ideal agents are never such that they know an event E, know that they know E, know that they know that they know E, ... *ad infinitum*.

The above argument shows that non-ideal agents are never such that there exists some point n for the event E such that every point $m > n$ is 1. Real, non-ideal agents may know that they don't know some event E, and they may know that they know *this*, and they may know that they know *this*. But, the above argument shows that such agents cannot extend this sequence infinitely (on pain of the existence of self-evident events). Now, my reasoning in conjunction with Lemma 5.1 is not enough to show that Unawareness of Knowledge is true—just because it is not the case that there exists some point n in the infinite iterative knowledge sequence such that every point $m > n$ is a 1, does not establish that that there is some (possibly different) n' such that every $m' > n'$ is a 0. Indeed, given what I've said so far, either of the following repeating sequences are possible:

(Seq. 5) ...101010101E

(Seq. 6) $\ldots 010101010E$.

Were either such sequence plausibly true of any real, non-ideal agent, it would provide a counterexample to Unawareness of Knowledge (as there would never be some point n such that every point $m > n$ is a 0). Could (Seq. 5) or (Seq. 6) ever be true of a real, non-ideal agent?

A third consideration for Unawareness of Knowledge concerns the difficulty of higher-order knowledge. (Seq. 5) and (Seq. 6) seem unattainable for real, non-ideal agents because they require an agent to have vast amounts of higher-order self-knowledge. Timothy Williamson (2000) gives a kind of analogy that is helpful on this issue. As he says, "the crude point is that iterating knowledge is hard, and each iteration adds a layer of difficulty".[28] Roughly, to know an event E or a proposition p is to have the resources to be able to discriminate between the world's being E or not-E (given one's knowledge or evidence). But to know that one knows is to have the resources to be able to discriminate between the world's being such that one knows E and one does not know E, and such discriminatory ability seems harder to achieve because it involves discriminating between knowledge *and* whether E. Higher-order knowledge is even more demanding. The idea is that eventually all real, non-ideal agents run out of the appropriate cognitive resources and fail to know such higher-order self-knowledge claims. This consideration may be thought of as a generalization of the first consideration (real agents do not have the cognitive resources to consider or entertain every point on the infinite iterative knowledge sequence).

In general, sequences that pose potential counterexamples to Unawareness of Knowledge are repeating sequences with some repeating positive element.[29] Beyond

[28]See Williamson (2000), p. 17 and p. 122.

[29]There are other potential iterative knowledge sequences that could work as counterexamples to Unawareness of Knowledge. For instance, consider some iterative knowledge sequence generated by choosing a "0" or "1" randomly at each point. Such a process could generate the following:

the three above considerations, such sequences (if true of an agent) are "unstable" interpretively—that is, it is hard to understand an agent or an agent's epistemic situation (what she knows and how she knows it) such that she is able to satisfy such sequences. Consider the first three points of (Seq. 5):

$101E$.

If these three points are true for some agent S, then the following is also true of S: "$K(E) \wedge \neg K(K(E)) \wedge K(\neg K(K(E)))$". While such a claim is not contradictory, it is difficult to understand an agent that satisfies these conditions. For instance, what story might we tell for why the second point or the second conjunct "$\neg K(K(E))$" is true? Perhaps, S does not know that she knows E because she is not paying attention to what she knows. It is then hard to reconcile this interpretation with the third point or the the third conjunct, namely that S knows that she does not know that she knows E. If S is not paying attention to what she knows about E then how does she know that she does not know that she knows E? Perhaps, instead, S does not know that she knows E because she has no information or evidence about whether she knows E. But, again, it is then hard to reconcile this interpretation with the fact that she knows that she does not know that she knows E. If she knows that she does not know that she knows E, it seems that she has *some* information or evidence about her higher-order knowledge of E. How, then, does she not know that she knows E?

Again, I want to stress that there is no contradiction in the proposition "$K(E) \wedge \neg K(K(E)) \wedge K(\neg K(K(E)))$". But there is an oddness. To understand

(Seq. 7) $\ldots 011001010010E$.

Again, were such a sequence true for some non-ideal agent, it would provide a counterexample to Unawareness of Knowledge. But the considerations above work against there being such a true sequence for a non-ideal agent. Real, non-ideal agents do not have the cognitive resources to satisfy such a knowledge claim.

this oddness, consider a case where an agent attempts to make clear her lack of knowledge. Consider a case where an agent utters the following:

I know E but I don't know that I know E.

Such an utterance is semantically odd, though it is perhaps also not contradictory. Such semantic oddness also seems to stem from a kind of epistemic instability—if S utters the sentence 'I know E', it is then hard to understand how S does not thereby know that she knows E.[30] This semantic oddness provides a model for the "informational" or "evidential" oddness from repeating sequences such as (Seq. 5) and (Seq. 6). From (Seq. 5), if one has the information or evidence that E, and one has the information or evidence that one does not know that one knows E, it is hard to understand why S does not have the information or evidence that she knows E. There are certainly stories that can be told. But the storytelling is more difficult given that (Seq. 5) is *infinite*. If (Seq. 5) is true for some agent S, then S always has information or evidence for the odd numbered points in the sequence, but S does not have information or evidence for the even numbered points. Such an agent would systematically not know that she knows some higher-order knowledge claim, but would systematically know that she knows *this* (which is, of course, a higher-order knowledge claim). I cannot imagine a plausible explanation for why an agent might have such a systematic deficit in knowledge.

Repeated sequences such as (Seq. 5) and (Seq. 6) do not seem plausibly true for real, non-ideal agents. In conjunction with the above considerations, I suggest that Unawareness of Knowledge is true. Hence, we are always unaware to some degree of what we know. The epistemic peculiarity is, of course, that we now know and are aware that we are always unaware to some degree of what we know.

[30]However, by Lemma 5.1 it must be the case that such higher-order knowledge eventually runs out. Certainly, due to time and memory constraints, an agent cannot utter every member from the corresponding infinite iterative knowledge sequence.

REFERENCES

Aumann, R. (1976). Agreeing To Disagree. *The Annals of Statistics*, **4**(6), pp. 1236–1239.

Bartsch, K. and H. Wellman (1995). *Children Talk About the Mind*. Oxford University Press.

Binmore, K. (1992). *Fun and Games: A Text on Game Theory*. D.C. Heath and Company.

Bishop, M. and J. Trout (2005). *Epistemology and the Psychology of Human Judgment*. Oxford University Press.

Carroll, L. (1958). *Symbolic Logic and the Game of Logic*. Dover.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

Chwe, M. (2001). *Rational Ritual: Culture, Coordination, and Common Knowledge*. Princeton University Press.

Cubitt, R. and R. Sugden (2003). Common Knowledge, Salience and Convention: A Reconstruction of David Lewis' Game Theory. *Economics and Philosophy*, **19**, pp. 175–210.

Dekel, E., B. L. Lipman, and A. Rustichini (1998). Standard State-Space Models Preclude Unawareness. *Econometrica*, **66**(1), pp. 159–173.

Doyle, A. C. (1901). The Adventure of Silver Blaze. *The Strand Magazine*.

Dretske, F. (1970). Epistemic Operators. *The Journal of Philosophy*, **67**, pp. 1007–1023.

Dretske, F. (2005). *Contemporary Debates in Epistemology*, chapter The Case Against Closure, pp. 13–26. Blackwell.

Ernst, Z. (2012). Convention and Bounded Rationality. Draft.

Fagin, R. and J. Y. Halpern (1988). Belief, Awareness, and Limited Reasoning. *Artificial Intelligence*, **34**, pp. 39–76.

Fagin, R., J. Y. Halpern, Y. Moses, and M. Y. Vardi (1995). *Reasoning About Knowledge*. MIT Press.

Fantl, J. and M. McGrath (2009). *Knowledge In an Uncertain World*. Oxford University Press.

Frigg, R. and S. Hartmann (2012). Models in Science. *Stanford Encyclopedia of Philosophy*.

Fundenberg, D. and J. Tirole (1991). *Game Theory*. MIT Press.

Geanakoplos, J. (1992). Common Knowledge. *The Journal of Economic Perspectives*, **6**(4), pp. 53–82.

Goldman, A. I. (1976). Discrimination and Perceptual Knowledge. *The Journal of Philosophy*, **73**(20), pp. 771–791.

Gray, J. (1978). Notes on Data Base Operating Systems. *Lecture Notes in Computer Science*, **60**, pp. 393–481.

Halpern, J. Y. and Y. Moses (1990). Knowledge and Common Knowledge in a Distributed Environment. *Journal of the Association for Computing Machinery*, **37**(3), pp. 549–587.

Hawthorne, J. (2004). *Knowledge and Lotteries*. Oxford University Press.

Hawthorne, J. (2005). *Contemporary Debates in Epistemology*, chapter The Case For Closure, pp. 26–43. Blackwell.

Hawthorne, J. and J. Stanley (2008). Knowledge and Action. *The Journal of Philosophy*, **105**(10), pp. 571–590.

Heal, J. (1978). Common Knowledge. *The Philosophical Quarterly*, **28**(111), pp. 116–131.

Hintikka, J. (1962). *Knowledge and Belief*. Cornell University Press.

Huemer, M. (2007). Epistemic Possibility. *Synthese*, **156**, pp. 119–142.

Kahneman, D. and A. Tversky (1973). On the Psychology of Prediction. *Psychological Review*, **80**, pp. 237–251.

Kvanvig, J. L. (2006). Closure Principles. *Philosophy Compass*, **1**(3), pp. 256–267.

Lewis, D. (1969). *Convention*. Harvard University Press.

Lewis, D. (1978). Truth In Fiction. *American Philosophical Quarterly*, **15**, pp. 37–46.

Lewis, D. (1996). Elusive Knowledge. *Australasian Journal of Philosophy*, **74**(4), pp. 549–567.

Lipman, B. L. (1995). Information Processing and Bounded Rationality: A Survey. *The Canadian Journal of Economics*, **28**(1), pp. 42–67.

McDowell, J. (1989). One Strand in the Private Language Argument. *Grazer Philosophische Studien*, **33/34**, pp. 285–303.

Moore, G. E. (1962). *Philosophical Papers*. Collier Books.

Nozick, R. (2001). *Invariances: The Structure of the Objective World*. Harvard University Press.

Osborne, M. J. and A. Rubinstein (1994). *A Course in Game Theory*. MIT Press.

Paternotte, C. (2011). Being Realistic About Common Knowledge: A Lewisian Approach. *Synthese*, **183**(2), pp. 249–276.

Pollock, J. L. and J. Cruz (1999). *Contemporary Theories of Knowledge*. Roman and Littlefield.

Rubinstein, A. (1989). The Electronic Mail Game: Strategic Behavior Under 'Almost Common Knowledge'. *The American Economic Review*, **79**(3), pp. 385–391.

Rubinstein, A. (1998). *Modeling Bounded Rationality*. MIT Press.

Schelling, T. (1960). *The Strategy of Conflict*. Harvard University Press.

Sharon, A. and L. Spectre (2008). Mr. Magoo's Mistake. *Philosophical Studies*, **139**, pp. 289–306.

Simons, D. and C. Chabris (1999). Gorillas in our Midst: Sustained Inattentional Blindness for Dynamic Events. *Perception*, **28**, pp. 1059–1074.

Sorensen, R. (1988). *Blindspots*. Oxford University Press.

Stalnaker, R. C. (1975). Indicative Conditionals. *Philosophia*, **5**(3), pp. 269–286.

Stalnaker, R. C. (1987). *Inquiry*. MIT Press.

Stalnaker, R. C. (1999). *Context and Content*. Oxford University Press.

Stanley, J. (2005). *Knowledge and Practical Interests*. Oxford University Press.

Stine, G. C. (1976). Skepticism, Relevant Alternatives, and Deductive Closure. *Philosophical Studies*, **29**, pp. 249–261.

Vanderschraaf, P. and G. Sillari (2009). Common Knowledge. *Stanford Encyclopedia of Philosophy*.

Von Wright, G. (1951). *An Essay in Modal Logic*. North-Holland Publishing Co.

Williamson, T. (2000). *Knowledge and its Limits*. Oxford University Press.

Williamson, T. (2011). *Evidentialism and its Discontents*, chapter Improbable Knowing. Oxford University Press.