

Epistemic paradox and explicit modal logic

by

Walter Dean

A dissertation submitted to the Graduate Faculty in Computer Science in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

2010

©2010

Walter Dean

All Rights Reserved

This manuscript has been read and accepted for the Graduate Faculty in Computer Science in satisfaction of the dissertation requirements for the degree of Doctor of Philosophy.

Date

Sergei Artemov
Chair of Examining Committee

Date

Theodore Brown
Executive Officer

Melvin Fitting

Leon Horsten

Rohit Parikh

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

Epistemic paradox and explicit modal logic

by

Walter Dean

Advisor: Sergei Artemov

The dissertation presents a unified treatment of what I refer to as the epistemic paradoxes (i.e. the Knower Paradox of Montague and Kaplan [76] and Fitch’s Paradox of Knowability [27]) in the context of a quantified extension of Artemov’s [3] Logic of Proofs [LP]. The system adopted is a modified form of Fitting’s [32] Quantified Logic of Proofs [QLP] and contains not only so-called explicit modalities (i.e. statements of the form $t : \varphi$ with intended interpretation of “ t denotes a proof of φ ”) but also first-order quantifiers intended to range over proofs or informal justifications. This allows for a formalization of a justification-based notion of knowledge via statements of the form $(\exists x)x : \varphi$.

The first chapter seeks to motivate the use of explicit modalities for reasoning about knowledge and informal provability relative to the tradition of

epistemic logic. The second chapter contains a consolidated treatment of the syntax and arithmetic and semantics of LP and QLP. A relational semantics for QLP is developed and several new results are presented (valid and invalid forms of explicit Barcan and converse Barcan formulas). The third chapter focuses on the reconstruction of the Knowers Paradox in QLP. The central observation is that by analyzing the $K(x)$ predicate employed by Montague and Kaplan in terms of proof quantification by replacing $K(\ulcorner \varphi \urcorner)$ with $(\exists x) : \varphi$, we realize that an additional logic principle is required for the derivation of the Knowers. This principle – which is similar to a form of Universal Generalization for proofs anticipated by Gödel [37] – is examined both proof theoretically and through the use of an arithmetic semantics for QLP. The fourth chapter focuses on the reconstruction of the Knowability Paradox in QLP. The central observation is the so-called Knowability Principle on which the paradox is based – i.e. the claim that if φ is true, then it is possible to know φ – is most appropriately formulated via the use of proof quantifiers as $(\exists x)x : \varphi \rightarrow \varphi$ as opposed to the use of propositional possibility and knowledge operators.

Acknowledgements

I dedicate this dissertation to my parents in gratitude for all their support.

Foremost in any list of acknowledgments I might devise for this work must appear the name of my friend and collaborator Hidenori Kurokawa. This work would not exist in its present form were it not for him. I also wish to acknowledge the support of my advisor, Sergei Artemov. He not only made it possible for me to write this work, but more importantly, he provided me with an intellectual home at the Graduate Center. I have profited greatly from years of stimulating conversation with him, and it goes without saying that his work forms the foundation on which this work is based. I was very fortunate that the complement of my committee was equally outstanding. Leon Horsten stepped in as my external committee member and provided many perceptive comments and questions. Melvin Fitting and Rohit Parikh have taught me much throughout the years. I look forward to continuing conversations with all of them. And finally, my time at CUNY would not have

been as pleasurable or productive without the many colleagues and friends I have been fortunate enough to have. I wish them to know how much I have enjoyed our conversations on topics both intellectual and ephemeral. I hope they will forgive my adoption (once again) of the following device: Thank you to Evangelia Antonakos, Yegor Bryukhov, Joseph Driscoll, David Etlin, Jeanne Funk, Evan Goris, Karen Kletter, Roman Kuznets, Florian Lengyel, Natasha Novak, Eric Pacuit, Frank Pupa, Bryan Renne, David Pereplyotchik, Dustin Mulcahey, Susan Schweitzer, Terence Swaine, Noson Yanofsky, Stathis Zachos.

Contents

1	Introduction	1
2	Explicit modal logic	28
2.1	Background	28
2.2	The Logic of Proofs	40
2.2.1	The language of LP	40
2.2.2	Axiomatization	42
2.2.3	Basic results	43
2.2.4	Semantics	47
2.3	The Quantified Logic of Proofs	52
2.3.1	The language of QLP	53
2.3.2	Axiomatization	54
2.3.3	Semantics	65
2.3.4	Relationship with S4	85
2.3.5	Barcan- and Converse Barcan-like formulas	86

3	QLP and the Paradox of the Knower	94
3.1	Introduction	94
3.2	The reconstruction of the Knower in modal logic	122
3.3	The reconstruction of the Knower in QLP	137
3.4	Arithmetic Semantics for QLP	144
3.5	GJUG, consistency and internalization	163
4	QLP and the Knowability Paradox	183
4.1	Introduction	183
4.1.1	A proof-based approach to the Paradox	188
4.1.2	Knowledge, proof and explicit modal logic	194
4.2	Neo-verificationism, knowledge and proof	203
4.3	Reconstruction and consistency	221
4.3.1	QLPE and the reconstruction of the Fitch derivation . .	222
4.3.2	The consistency of (KP4) and (NonOm3)	236
4.4	Comparisons and contrasts	244
	Bibliography	256

Chapter 1

Introduction

What I will refer to herein as the *epistemic paradoxes* are two derivations in formal logic which reflect tensions between the properties which can hold of an operator or predicate K with the intended interpretation “knows that.” These antimonies (or purported antinomies) are conventionally referred to as the Paradox of the Knower and the Paradox of Knowability. Both were introduced in the early 1960s, the former by Richard Montague and David Kaplan [76] and the latter by Frederic Fitch [27].

Although the two paradoxes came to light at about the same time, they engage with the everyday concept of knowledge in very different ways. One of the simplest formulations of a Knower-like Paradox¹ is as follows: if a predicate $K(x)$ (intended to represent that the sentence with godel-number x is known) is adjoined to a arithmetic theory T like first-order Peano Arithmetic [PA] which is sufficiently strong to encode operations on its own syntax and

it is also assumed that knowledge implies truth (i.e. $T \vdash K(\ulcorner \varphi \urcorner) \rightarrow \varphi$) and that provability in T implies knowledge (i.e. if $T \vdash \varphi$, then $T \vdash K(\ulcorner \varphi \urcorner)$), then T is inconsistent. The derivation of the contradiction passes through the fact that T is sufficiently strong to prove the existence of self-referential statements about $K(x)$ such as

$$(1.1) \quad \delta \leftrightarrow \neg K(\ulcorner \delta \urcorner)$$

This sentence receives the intended interpretation *δ is true iff δ is not known*. As such, the Knower can be taken to reflect a potential conflict between certain intuitively evident principles about knowledge and the possibility of accommodating self-reference in a single formal theory.²

The Knowability Paradox is perhaps best understood not as an out-and-out contradiction derived from indisputable premises, but rather as a strikingly counter-intuitive consequence of a view which does not obviously appear to be logically incoherent. The view in question is a theory about truth known as *verificationism* according to which no true statement is beyond potential verification or, as it is sometimes put, all true propositions are pos-

²The result cited here is commonly referred to as Montague's Paradox and has a somewhat simpler formulation than the result originally presented as "The Knower" by Montague and Kaplan. It was first explicitly derived by Montague [75] where it was presented as antinomy concerning the notion of necessity. However, its premises remain plausible when $K(x)$ is taken to express knowledge and in Chapter 2 I will argue that it is this latter result which rests on more fundamental epistemic principles.

sibly known – a statement referred to as the *Knowability Principle* [(KP)]. According to most commentators on the Knowability Paradox, this view may be formulated in a modal language containing propositional operators for knowledge (K) and possibly ($\Diamond$) via the following principle:

(KP1) for all φ , $\varphi \rightarrow \Diamond K\varphi$

This principle has been widely taken to be a schematic representation of the so-called anti-realist view of truth favored by philosophers such as Michael Dummett (cf., e.g., [22]). The Knowability Paradox corresponds to the observation that when we also take as a premise the presumably uncontentious claim that there are unknown truths (formulated as $\psi \wedge \neg K\psi$ for some ψ) then it is possible to derive $\varphi \rightarrow K\varphi$ for all φ assuming only very weak principles about K and $\Diamond$. As this latter statement can be taken to express that all truths are known – a claim which appears to be straightforwardly false – the Knowability Paradox is often presented as constituting a refutation of verificationism in the form of a reductio.

The epistemic paradoxes thus have quite different intellectual origins and we will see that their subsequent investigation has touched on a variety of very different topics in philosophical and mathematical logic. However the results of Fitch and Montague and Kaplan also serve as landmarks in the

early uses of formal logic to facilitate reasoning about knowledge. This is true in the narrow sense that Montague’s original work helped to delimit the range of options for how provability could be treated as a modality. This initiated work which ultimately lead to the isolation of the Gödel-Löb logic GL as the provability logic of PA. But it also true in the broader sense that Fitch’s treatment of ‘it is known that’ as a propositional operator represents one of the earliest attempts to apply modal logic to facilitate reasoning about knowledge as a sentential operator, a point of view which is indirectly supported by some of the conclusions framed later by Montague [75]. Thus despite the fact that the epistemic paradoxes appear to be limitative in nature, they can also be taken as important early contribution to epistemic logic in the modal tradition of Hintikka [53] and von Wright [109].

The primary goal of this work will be to reconstruct and evaluate the significance of the epistemic paradoxes using tools from a family of systems known as *explicit modal logics*. Such systems share a common origin in Artemov’s [3] Logic of Proofs LP and can be taken to refine the systems in which the paradoxes were initially formulated by the use of so-called *explicit modalities*. Putting aside for the moment the distinction between the operator and predicate approaches to modality, we have seen that the original formulations of both the Knowability and Knower Paradoxes employed a symbol K

(applicable respectively sentences and names of sentences) with the intended interpretation

$$(1.2) \quad K\varphi \rightleftharpoons \varphi \text{ is known}$$

A language containing such a device allows us to formulate certain features of the everyday concept of knowledge such as the fact that knowledge of φ entails its truth – i.e.

$$(T) \quad K\varphi \rightarrow \varphi$$

– or that knowledge of φ entails knowledge of the fact that φ is known – i.e.

$$(4) \quad K\varphi \rightarrow KK\varphi$$

But at least at a syntactic level, systems of propositional modal logic do not seek to analyze the *meaning* of knowledge assertions in the sense of providing an account of what is required for a statement of the form $K\varphi$ to hold in a given situation.

In the context of traditional epistemic logic, the epistemic paradoxes might be taken to show that conflicts already arise among principles about knowledge which can be stated at the propositional level. But since these principles include such apparent truisms as the claim that knowledge entails truth (i.e. $K\varphi \rightarrow \varphi$) or that knowledge distributes over conjunction (i.e.

$K(\varphi \rightarrow \psi) \rightarrow (K\varphi \rightarrow K\psi)$) they might even be taken to show that our everyday intuitions about knowledge are defective or even potentially incoherent. One of the aims of the present work will be to argue for the thesis that rather than reflecting conflicts within our everyday intuitions, the epistemic paradoxes are consequences of the expressive limitations of the formal systems in which they were originally formulated. I will argue that this is true in both the instrumental sense that the adoption of a more expressive language allows for a resolution to the paradoxes which is compatible with everyday intuitions, and also in a more historical sense in that the languages employed in the original formulation are too impoverished to express essential features of the relevant epistemic concepts which are central to the intellectual contexts in which the paradoxes were originally framed.

The most fundamental feature of knowledge which I will argue is left out of the traditional formulation of the epistemic paradoxes is that of proof or justification. In order to see why this might be significant, we may begin by recalling the traditional analysis of knowledge as justified, true belief. According to this analysis, statements of the form *agent i knows φ* ought to be analyzed as follows:

$$(1.3) \quad \text{i) } \varphi \text{ is true}$$

ii) i believes φ

iii) i 's belief in φ is justified

It need hardly be mentioned that although (1.3) was for a long time held to express something like a conceptual analysis of the everyday conception of knowledge, its status is now controversial within mainstream epistemology. The basis of this controversy derives in large part to the familiar observation (commonly attributed to Gettier [34]) that there exist situations in which i may possess justification in virtue of which he believes some true statement φ , but still does not know φ . This has lead some theorists (e.g. Goldman [39]) to seek after an appropriate fourth condition to add to the justified-true-belief [JTB] account. However, the task of locating such a uniform condition of this sort has proven surpassingly difficult and at present there are few, if any, proposals have which escaped substantial criticism within mainstream epistemology (cf., e.g., [96] for a survey). Meanwhile other theorists – the so-called *contextualists* (e.g. Lewis [66]) – have claimed that what ought to count as sufficient for φ varies with the context in which we wish to attribute knowledge of φ . And others – the so-called *reliabilists* (e.g. Plantinga [80]) – have proposed that i 's knowledge of φ requires only that he has acquired his belief in φ by some reliable mechanism and that epistemology therefore

ought to stay out of the business of analyzing the notion of justification itself.

Such trends notwithstanding, I will argue that the justificatory component of the traditional analysis of knowledge is central to the intellectual setting of both the epistemic paradoxes. Although a thorough accounting of the considerations on which this claim is based would be premature here, its basis is quite apparent in the case of the Knowability Paradox. The fundamental observation is that although some verificationists (most famously Dummett in [22]) have attempted to frame their views about truth in terms of the concepts of *knowledge* and *possibility*, it is much more common for them to do so using the concepts of *proof* and *existence*. In particular, statements of principles resembling the following

(KP2) if φ is true, then there exists a proof of φ

can be found throughout the writings not only of Dummett, but also the other two prominent contemporary verificationists Dag Prawitz and Per Martin-Löf.³ I will argue below that their endorsement of such a principle stems not from a desire to define truth in terms of knowledge when this latter notion

³As I will discuss further in Chapter 4, the use of term “verificationist” to describe the views of these theorists is something of an anachronism as their views derive more closely from Brouwer (in the case of Dummett) and Gentzen (in the case of Prawitz and Martin-Löf) than they do from the form of verificationism associated with the Vienna Circle. As such, I will adopt the term *neo-verificationist* to describe the contemporary theorists against whom the Knowability Paradox is most often targeted.

is understood in terms of our everyday conception. Rather, it derives from their background commitments to intuitionistic logic and mathematics. In this setting, the concept of truth is commonly explicated in terms of terms of provability in an appropriately constructive system. As such, I will argue that any formulation of the Knowability Paradox which seeks to impugn verificationism must be based on a formulation of (KP) which resembles (KP2) rather than (KP1).

This observation also suggests that in order to be faithful to the intentions of the neo-verificationists (and also to understand the claims of their critics), a formalization of (KP) must make explicit the relationship between knowledge and proof. In the context of the everyday conception of knowledge, this turns out to be a difficult task to achieve because (the contentions of the verificationists notwithstanding) it is not at all clear whether there is any accepted notion of proof which can sensibly be used to delimit the class of statements which we take ourselves to know. But at least in cases where φ is a mathematical or logical statement, it seems plausible to identify a justification for φ with a mathematical proof. Such a proof may take either the form of a derivation in an axiomatic system such as PA or ZF, or it might take the form of an informal mathematical argument, of the sort standardly presented in lectures and textbooks. On either understanding, however, to say that i

knows φ appears to require that i is in possession of a proof of φ . It thus follows that to say that a logical or mathematical statement φ is *possibly known* by i is presumably to state that a proof of φ exists where the sense of existence in question allows that there might exist a proof which i has not yet himself constructed (or learned about through other sources). These observations suggest that we formulate the epistemic concepts appearing in the premises of the Knowability Paradox as follows:

(1.4) a) φ is knowable $\Leftrightarrow$ there exists a proof of φ

b) φ is known $\Leftrightarrow$ there exists a constructed proof of φ

Below I will argue that (1.4a,b) come substantially closer to expressing the intentions of the verificationists when they appear to endorse statements resembling (KP). On this basis, I will also suggest that substantial insight can be gained into how the Knowability Paradox can be defused. To reiterate, the central aspect of the reformulation I will recommend is that rather than treating knowledge as an unanalyzed notion which we then proceed to reason about propositionally, it is also possible to expose the tactic quantification over justifications which, at least according to the JTB account, is implicit in the right-hand sides of (1.4a,b). Once this step is taken, however, we must immediately confront the question of how these analyses can be appropriately

formalized in order to reconstruct the reasoning of the Knowability Paradox. It is the need for doing so which motivates the application of explicit modal logic to the epistemic paradoxes.

Rather than propositional modal operators, systems in the family of the Logic of Proofs contain so-called *explicit modalities* of the syntactic form $t : \varphi$, wherein t is itself a potentially structured entity known as a *proof term*. Such statements have the intended interpretation

$$(1.5) \quad t : \varphi \Rightarrow t \text{ is evidence for } \varphi$$

The essential feature of explicit modalities is thus that they are intended to express an evidentiary relation which obtains between a proposition φ and some sort of justificatory object denoted by t . Such evidence might take the form of what I will refer to as *everyday evidence or justification* – i.e. considerations such as sense data, testimony, etc. which we commonly take to form the basis of most of our empirical knowledge. However, the interpretation which will be most relevant in the sequel is one on which proof terms are taken to denote mathematical proofs. And as I have already noted, this notion itself splits into formal and informal variants.

The intention of most early work on LP was to interpret proof terms as denoting axiomatic proofs in a formal system such as PA and the relation

expressed by $:$ as corresponding to that which holds between such a proof and the theorem which it demonstrates. This idea can be made precise by defining a mapping $(\cdot)^*$ which maps formulas of **LP** to gödel numbers of arithmetic formulas and proof terms to gödel numbers of arithmetic proofs relative to a proof predicate $\text{Prf}(x, y)$ for (e.g.) **PA**. If we then adopt the following interpretation of explicit modalities

$$(1.6) \quad (t : \varphi)^* \iff \text{Prf}(\ulcorner t^* \urcorner, \ulcorner \varphi^* \urcorner)$$

it may be shown that the image under $(\cdot)$ of each theorem of **LP** is not only true in the standard model of arithmetic, but also provable in **PA**.⁴

The foregoing observations suggest that it is at least coherent to develop a calculus for reasoning about statements of the form $t : \varphi$ under the interpretation given by (1.5). In fact, once we have adopted a language containing explicit modalities, it becomes possible to express a variety of principles which are intuitively valid not just for formal proofs, but also possibly for informal proofs and everyday justifications as well. Among these are the following schemas which correspond to the axioms of **LP**:

$$(1.7) \quad \text{a) } t : \varphi \rightarrow \varphi \qquad \qquad \qquad (\text{“reflection”})$$

⁴This result requires that the proof predicate $\text{Prf}(x, y)$ satisfy some properties which go modestly beyond the standard Hilbert-Bernays conditions. A precise statement will be given in Chapter 2.

- b) $t : (\varphi \rightarrow \psi) \rightarrow (s : \varphi \rightarrow t \cdot s : \psi)$ (“application”)
- c) $t : \varphi \rightarrow !t : t : \varphi$ (“proof checker”)
- d) $t : \varphi \rightarrow t + s : \varphi$ (“sum”)
- e) $\vdash c : \varphi$ where φ is a propositional tautology or an instance of a) -
 d) and c is a proof constant
 (“axiom necessitation”)

These statements demonstrate the sense in which proof terms can be structured entities in the sense of being built up inductively from so-called *proof variables* and *proof constants* using operations denoted by $\cdot$, $!$ and $+$. As we will see in Chapter 2, each of these elements can be given a precise arithmetic interpretation. But subject to the following glosses, the LP axioms themselves can be taken to provide implicit definitions:

- (1.7') a) If t is evidence for φ , then φ is true.
- b) If t is evidence for $\varphi \rightarrow \psi$ and s is evidence for φ , then evidence for ψ can be obtained by combining t and s .
- c) If t is evidence for φ , then $!t$ is evidence for $t : \varphi$
- d) If t is evidence for φ , then the result of conjoining t with additional evidence s is still evidence for φ

(1.7'a) expresses what I refer to as the *factivity* of the justification denoted by t – i.e. that if t is evidence for φ , then φ is true. It is evident that this is *not* a feature of the everyday concept of justification. For instance, per Descartes, although my perception t of a computer on my desk justifies my belief in the proposition φ that there is indeed a computer on my desk, I could possess t and yet φ still might be false – I might be dreaming or the victim of a evil demon's deception. However, factivity presumably *is* a central feature of our concept of informal mathematical proof. This is to say that it is presumably part of what we mean when we say that t is a proof (as opposed to merely an apparent proof) of φ , that φ must be true.

There are, of course, well known reasons why it may not be possible to carry over this property of intuitive notion of proof to our formal characterization of what it is to be a mathematical proof. Many of these relate to Gödel's Second Incompleteness Theorem and will be discussed in detail in Chapter 3. However, it is important to realize that if we confine attention to proofs which can be denoted by (standard) numerals – a feature which will be secured by the definition of the arithmetic interpretation $(\cdot)^*$ described above – then it may also be shown that (1.7a) will hold under this interpretation.⁵

⁵Again, more details will need to be supplied below to make this claim precise. But the essential idea is that $t^* = n \in \mathbb{N}$ then $(t : \varphi)^* = \text{Prf}(\bar{n}, \ulcorner \varphi \urcorner)$ will then be a Δ_1 in the language of arithmetic. This statement will then be provable or refutable in PA, according as to whether φ is true in the standard model N . In either case, it follows that

(1.7'b) expresses a principle which appears to be evident for not only the everyday notion of justification, but also for the concepts of both informal and formal proof. First, on both the everyday and informal proof construal, suppose that t is evidence for the material conditional $\varphi \rightarrow \psi$ and that s is evidence for φ . Then we can presumably come to be justified in believing ψ by combining s and t in a manner that mimics reasoning by detachment. In other words, we can convince ourselves that ψ is true by first surveying t on the basis of which we have evidence to assert $\varphi \rightarrow \psi$ and then surveying s . By combining t and s together and then detaching, we are thereby justified in concluding ψ . An informal explanation of this of course leaves open how the operation on justifications denoted by $\cdot$ ought to be interpreted. However, In the case of formal proofs, it appears to make sense to speak of “combining” justifications as concatenation of derivations and of detachment as the performance of modus ponens within the axiomatic system. On this basis, it possible to provide a straightforward arithmetic interpretation of $\cdot$ as an operation on the godel number of proofs. However, if our aim is simply to reason about the relation given by (1.5) itself, then nothing prevents us from simply viewing (1.7'b) as providing an implicit definition of the properties which such an operation must possess.

$(t : \varphi \rightarrow \varphi)^* = \text{Prf}(\bar{n}, \ulcorner \varphi \urcorner) \rightarrow \varphi$ is both true in N and provable in PA.

Similar considerations can be adduced in favor of adopting (1.7'c) as an axiom describing both everyday justifications and also formal and informal proofs. For note read according to (1.5), this statement expresses that if t denotes a justification of φ , then there is another justification denoted by $!t$ which serves as evidence of this former fact. This is true on the formal provability interpretation due to the well-known ability of systems like **PA** to internalize reasoning about their own proofs – i.e. if $\text{PA} \vdash \text{Prf}(\bar{n}, \overline{\ulcorner \varphi \urcorner})$, then the derivation which witnesses this fact can itself be internalized to yield a proof of $\text{Prf}(\bar{m}, \overline{\ulcorner \text{Prf}(\bar{n}, \overline{\ulcorner \varphi \urcorner}) \urcorner})$ for some m which can be uniformly computed from n . As long as we assume that informal proofs are themselves objects which we can reason about within mathematics, we can use this result as the basis for a heuristic argument that (1.7'c) also ought to hold for the informal provability interpretation – i.e. if t denotes an informal proof of φ , then it ought to be possible for us to recognize this fact by inspecting the structure of t . But such an operation can itself be taken as an informal proof of $t : \varphi$ and can hence be taken as a characterization of the meaning of $!$ operator.⁶

⁶Although this will not turn out to be relevant in the sequel, it bears noting that it is much less clear whether (1.7'c) when proof terms are taken to denote everyday justifications. In particular, the sorts of evidentiary considerations t which might be taken to justify φ are sufficiently heterogeneous that it is not even clear what it would mean to “internalize” such considerations in a manner which would bring one to have evidence for the claim that t was evidence for φ . For a different sort of argument against the implicit form of (1.7'c) – i.e. the so-called “K-K principle” in the form of 4 – see Williamson [117].

Finally it will be noted that (1.7'd) corresponds to a sort of monotonicity principle for evidence – i.e. it states that if t serves as evidence for φ then for any additional justification s , the result of considering t together with s is still a justification for φ . Such a principle is again notoriously *false* for the everyday concept of justification. For instance, I may take my evidence t that Nixon is a Quaker to justify the proposition that he is a pacifist. However, suppose s is additional evidence that Nixon is a Republican. It may well be that when I consider t and s as a conjoint body of evidence, I no longer think that together they serve to justify φ . However, the possibility of non-monotonicity seems to be absent from our concept of informal mathematical proof. For if t denotes a proof of φ , then per the factivity of this relation, φ is true from which it presumably follows that no other proof which I might learn about can interfere with this fact. If suitable properties are assumed about $\text{Prf}(x, y)$ (which among other things codify the fact that we are speaking of *multi-conclusion* proofs), then a formal analog of this can also be shown to hold.

The upshot of the foregoing considerations is that **LP** can be reasonably described as a formal calculus for reasoning about the relation expressed on (1.5) on at least the formal and informal proof construals of justification. As we will see below, this system is also closely related to the propositional

modal system **S4** and consequently to intuitionistic propositional calculus [IPC] via Gödel's [35] embedding of IPC into **S4**. This embedding is often presented as a kind of analysis of the so-called Brouwer-Heyting-Kolmogorov [BHK] semantics for intuitionistic logic as there is also a tradition – about which I will say more in Chapter 2 – of attempting to interpret the $\Box$ of **S4** as an informal provability operator. Hence there is already a precedent for using explicit modal logic to analyze the notion of constructive proof on which the verificationists have based their understanding of (KP).

A related feature of **LP** is its ability to internalize its own proofs in a manner reminiscent of the necessitation rule of **S4**. Recall that like all normal modal logics, **S4** contains a necessitation rule of the form

(Necessitation) If $\vdash \varphi$, then $\vdash \Box\varphi$.

On the informal provability interpretation, this principle is justified by the observation that whatever schematic principles about informal provability are provable in the system we adopt for reasoning about informal provability are, *ipso facto*, themselves informally provable. A related phenomena obtains for **LP** and related systems, the simplest form of which is as follows

(Constructive Necessitation Theorem) If $\mathbf{LP} \vdash \varphi$, then $\mathbf{LP} \vdash t : \varphi$ for some proof term t .

A paradigmatic feature of explicit modal logics is thus that they are able to internalize their own proofs in the sense that if φ is provable in such a system, it is possible to construct a proof term which encodes its derivation such that $t : \varphi$ is itself provable in the system.

I will argue below that internalization is an essential ingredient of the Knower Paradox. This feature is obscured by the derivation of the paradox originally given by Montague and Kaplan wherein no direct analog of Necessitation is assumed. However, Montague and Kaplan do assume the following the principle

$$(U) \quad K(\ulcorner K(\ulcorner \varphi \urcorner) \rightarrow \varphi \urcorner)$$

which is readily recognized as a necessitated form of the modal T (or “reflection”) axiom. I will argue that there are good grounds for taking the version of the Paradox given above (in which Necessitation is assumed rather than U) as the most fundamental form of the result. However, in the context of either the formal or informal provability interpretations of justifications, we have just seen that there is no reason to regard Necessitation as a basic rule. Rather if the system in question derives φ , it is reasonable to expect that we ought to be able to construct an object t which represents this derivation and derive the fact that t justifies φ within the system itself. However, as

we will see below, internalization of this sort requires only the necessitation of axioms (as provided for in **LP** by (1.7e)) and not theorems in general.

The one component missing from the basic system **LP** which is required to reconstruct the epistemic paradoxes is an ability to quantify over proofs in the object language. Such a capacity is obviously required to formalize (KP2). And I will also propose below that considerations related to the role of internalization suggest that the concept of knowledge which figures in the Knower Paradox also suggests that the knowledge predicate $K(x)$ employed by Montague and Kaplan should be analyzed according to (1.4a). This is to say that rather than treating knowledge as a predicate of (the godel numbers) of sentences, the context of the paradox suggests that knowledge of φ ought to be analyzed in terms of the existence of a proof of φ .

As we will see below, the language of **LP** is traditionally formulated so as to contain free proof variables. One motivation for subsuming such quantification into the object language is thus to be able derive appropriate generalizations of statements like

$$(1.8) \quad \begin{array}{l} \text{a) } x : (\varphi \rightarrow \psi) \rightarrow (y : \psi \rightarrow x \cdot y : \psi) \\ \text{b) } x : \varphi \rightarrow !x : x : \varphi \end{array}$$

These appear to express principles about informal proofs which remain valid when their free variables are understood as universally bound. On the basis of

this understanding, one might argue that the validity of statements involving proof quantifiers such as

$$(1.9) \quad \begin{array}{ll} \text{a) } (\forall x)(\forall y)[x : (\varphi \rightarrow \psi) \rightarrow (y : \psi \rightarrow x \cdot y : \psi)] \\ \text{b) } (\forall x)(\exists y)[x : \varphi \rightarrow y : x : \varphi] \end{array}$$

is already implicit in our understanding of explicit modalities.

With this said, it should also be kept in mind that variables play two very different roles in classical first-order logic. On the one hand, when bound by quantifiers, they serve as a semantic mechanism for expressing quantification over a domain of objects. But on the other hand, they also serve as a syntactic mechanism for displaying quantifier scope and binding relationship between argument places. One reflection of this is the Skolem Normal Form Theorem, an simple instance of which states that a first-order sentence $\forall x \exists y R(x, y)$ is satisfiable if and only if one of the form $\forall x R(x, f(x))$ is satisfiable (where f is a functional symbol in a possibly expanded language). Artemov [3] suggests that the Realization Theorem for LP can be taken as a partial analogy to such a result as it ensures that if $\vdash_{\mathbf{S4}} \varphi$, then $\vdash_{\mathbf{LP}} \varphi^r$ where r replaces occurrences of $\Box$ in negative position with free variables and positive occurrences of $\Box$ with functional combinations of these variables in a manner which exhibits the same sort of functional dependencies as a Skolem normal form.

The system I will adopt for reconstructing reasoning involving quantifica-

tion over proofs is a minor modification of the so-called Quantified Logic of Proofs proposed in Fitting [32] for which I retain the name QLP. It will contain formulas of the form $(\exists x)x : \varphi$ with the intended interpretation *there exists a proof of φ* . Using this construction it hence becomes possible to formalize (KP2) as

(KP3) for all φ , $\varphi \rightarrow (\exists x)x : \varphi$

and the fixed point statement used in the derivation of the Knower as

$$(1.10) \quad \delta \leftrightarrow (\neg \exists x)x : \delta$$

These reformulations will provide the starting points for my reconstructions of the epistemic paradoxes in Chapters 3 and 4.

As one might expect, the introduction of quantification into the basic language of LP introduces a number of both technical and conceptual complications to its interpretation. For on the one hand, devices must be found for binding proof variables in the context of (potentially nested) explicit modalities, performing substitution of proof variables by terms, and interpreting free proof variables. And on the other hand, it is reasonable to expect that as quantificational machinery is introduced into LP, it becomes possible to derive a variety of formulas which highlight the differences between formal

and informal provability. Paradigmatic amongst these is the generalized form of the LP explicit reflection axiom – i.e. the statement

$$(\mathsf{T}_q) \quad (\exists x)x : \varphi \rightarrow \varphi$$

I suggested above that there is a way of restricting the interpretation of proof terms so that the interpretation of the non-quantified form of this axiom is not only true (in the standard model) but also provable in PA. However, the natural arithmetic interpretation of T_q is

$$(1.11) \quad (\exists y)\text{Prf}(y, \overline{\ulcorner \varphi \urcorner}) \rightarrow \varphi$$

Although all instance of this principle are true in the standard model of arithmetic, it follows from the Second Incompleteness Theorem that certain instances (e.g. with $\varphi \equiv 0 = 1$) are not derivable in PA (as long as it is consistent). We also will see that the necessitated form of this statement – i.e.

$$(\mathsf{U}_q) \quad (\exists y)y : ((\exists x)x : \varphi \rightarrow \varphi)$$

– is also derivable in QLP. However, it is evident that there are arithmetic instances of the interpretation of (U_q) are not only unprovable in PA, but also *false* in the standard model. This is one of several reasons we will why the introduction of quantification into the language of LP introduces substantial interpretative challenges.

Fitting’s formulation of an elegant and flexible relational semantics for QLP may be taken as at least partial confirmation that such challenges can be met. However, I will suggest below that the decision to adopt quantification into the language of LP does not by itself provide us with completely determinate grounds for selecting one means of axiomatizing first-order reasoning about justifications over another. In particular, there are a number of ways the internalization of reasoning involving universal generalization might be handled between which our background intuitions about informal proof or justifications do not appear to decide. As I will explore further below, the different options appear to be connected with whether our concept of evidence is compatible with the existence of “merely potential” justifications or proofs – i.e. evidential items which need not exist in the actual state of affairs but which might exist in an alternative one.

The rest of this thesis will be structured as follows. In Chapter 2, I will present a unified exposition of LP and QLP. The LP portion will be largely expository in the sense that it will largely be recounting of traditional definitions and previously obtained results such as the Constructive Internalization and Realization Theorems mentioned above as well as the arithmetic soundness and completeness results for LP. My presentation of QLP will focus on different ways in which reasoning by Universal Generalization about proofs

can be internalized. Fitting has proposed that this can be accomplished in two different ways, respectively in [30] and [32], resulting into two systems which I will respectively call QLP_1 and QLP_2 . The system which I will refer to as QLP is a reformulation of QLP_2 which strengthens the so-called *Justified Universal Generalization* rule and drops and drops an axiom from the original system. This system will be shown to be sound and complete with respect to Fitting's [32] relational semantics for QLP_2 and also to have the same class of theorems as this system. The status of various analogs of the Barcan and Converse Barcan formulas will also be investigated.

Chapter 3 will be devoted to an exposition of several forms of the Knower Paradox, together with its reconstruction in QLP and discussion of its relationship to formal theories of truth and modality. This will provide a context in which to bring out the details of the argument sketched above for the implicit appeal to internalization in the derivation of Knower. I will then show how these derivations can be reconstructed in QLP by assuming appropriate self-referential statements as premises. I will argue that the locus of the paradox lies in the ability to internalize statements derived via the application of universal generalization over justifications. I will suggest that this refines and strengthens earlier analyses (e.g. [1], [25]) which called for the rejection of the necessitated form of the principle U mentioned above. This will in

turn provide a context in which to frame some additional results about QLP, such as the non-conservativity of various forms of justified universal generalization QLP and the possibility of formulating an arithmetic semantics for an appropriate subtheory.

Chapter 4 will be devoted to an exposition of the Knowability Paradox, together with its reconstruction in QLP and related reflections about its relationship to the formulation of verificationism using quantification over proofs. I will begin by surveying the views of the neo-verificationists in a manner which highlights their relationship to intuitionism and apparent acceptance of the principles like (KP2). Along the way I will adduce a conceptual argument that these theorists are already committed to the existence of non-constructed proofs. I will then show how both this argument and the original derivation of the paradox may be reconstructed in a version of QLP extended with a proof existence predicate $E(x)$. In light of the analysis of knowability in terms of proof existence suggested above, I will argue that the reconstructed derivation leads not to a paradoxical conclusion but instead provides another means of seeing how the verificationists are committed to non-constructed proofs. I will then turn to a formal demonstration of the consistency of (KP2) together with my proposed formalization of the existence of unknown truths.

Chapter 2

Explicit modal logic

2.1 Background

Both Artemov's Logic of Proof and Fitting's Quantified Logic of Proofs can be taken to have their intellectual origins in Gödel's [35] proposal that intuitionistic propositional calculus [IPC] can be interpreted using a propositional operator $\mathcal{B}$ with the intended interpretation

$$(2.1) \quad \mathcal{B}\varphi \quad \Rightarrow \quad \varphi \text{ is provable}$$

It will be recalled that intuitionistic truth is often explicated in terms of intuitionistic provability (cf., e.g., [105]). The interpretation of the intuitionistic connectives first proposed by Heyting [52] seeks to extend this characterization to arbitrary formulas by specifying so-called *proof conditions* which state the circumstances under which an object is to count as a proof of a complex statement in terms of the proof conditions of its constituents. Now enshrined

in the so-called Brouwer-Heyting-Kolmogorov interpretation of intuitionistic logic, these conditions include clauses such as the following:

(2.2) a) A proof of $\varphi \vee \psi$ is given by a proof of φ or a proof of ψ .

b) A proof of $\varphi \rightarrow \psi$ is a construction which permits us to transform any proof of φ into a proof of ψ .

Gödel defined the following embedding $(\cdot)^\circ$ from the language of IPC into the modal language containing the operator $\mathcal{B}$:

- $P^\circ = P$
- $(\varphi \wedge \psi)^\circ = \varphi^\circ \wedge \psi^\circ$
- $(\varphi \vee \psi)^\circ = \mathcal{B}\varphi^\circ \vee \mathcal{B}\psi^\circ$
- $(\varphi \rightarrow \psi)^\circ = \mathcal{B}\varphi^\circ \rightarrow \mathcal{B}\psi^\circ$
- $(\neg\varphi)^\circ = \neg\mathcal{B}\varphi^\circ$

Now consider the classical modal system **S4**:

Definition 2.1.1. A Hilbert system for the modal logic **S4** consists of the axioms

Taut all tautologies of classical propositional logic

$$(T) \quad \Box\varphi \rightarrow \varphi$$

$$(4) \quad \Box\varphi \rightarrow \Box\Box\varphi$$

$$(K) \quad \Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$$

together with Modus Ponens and the following rule (conventionally known as Necessitation)

(Nec) If $\vdash \varphi$, then $\vdash \Box\varphi$.

Gödel established the following

(2.3) If $IPC \vdash \varphi$, then $S4 \vdash (\varphi)^\circ$.

and conjectured the converse – i.e.

(2.4) If $S4 \vdash (\varphi)^\circ$, then $IPC \vdash \varphi$.

– which was later established by McKinsey and Tarski [73] using algebraic semantics.

Both Gödel and McKinsey and Tarski observed that an analogous result can be established relative to a variety of different embeddings.¹ Alternative

¹Gödel observed that changing the clauses for negation and conjunction to $(\neg\varphi)^\circ = \mathcal{B}\neg\mathcal{B}\varphi$ and $(\varphi \wedge \psi)^\circ = \mathcal{B}\varphi^\circ \wedge \mathcal{B}\psi^\circ$ also the same result to obtained. McKinsey and Tarski additionally showed that for the embedding $(\cdot)^-$ defined via $(P)^- = \mathcal{B}P$, $(\varphi \wedge \psi)^- = \varphi^- \wedge \psi^-$, $(\varphi \vee \psi)^- = \mathcal{B}\varphi^- \vee \mathcal{B}\psi^-$, $(\varphi \rightarrow \psi)^- = \mathcal{B}\varphi^- \rightarrow \mathcal{B}\psi^-$, $(\neg\varphi)^- = \neg\mathcal{B}\varphi^-$ we have that $S4 \vdash \varphi^- \leftrightarrow \mathcal{B}\varphi^\circ$.

embeddings based on both stronger and weaker modal logics are also known: **S4** can be replaced by **S4** + the Grzegorczyk axiom or by **S3** (Hacking [42]). Due to the non-uniqueness of both $(\cdot)^\circ$ and the choice of modal logic there does not seem to be a good argument for regarding this embedding as providing a meaning preserving *translation* of the BHK interpretation.² However, the embeddability of intuitionistic systems into modal systems in the vicinity of **S4** appears to be a robust phenomenon in general. For instance, Rasiowa and Sikorski [90] extended Gödel's result to intuitionistic predicate calculus (IPC) and Shapiro [94] and Goodman [41] have extended it to Heyting arithmetic [HA].³

Gödel also observed that the use of a modal operator with the intended interpretation (2.1) is subject to an important qualification. For suppose that we wish to interpret $\mathcal{B}\varphi$ as expressing provability in a particular formal system T containing a sufficient amount of arithmetic to formalize reasoning about its own syntax. In this case, we might expect that a connection between the operator and a provability predicate $Prov_T$ for T could be established via an interpretation which satisfied

²It does, however, appear that Gödel's choice of embedding was inspired by an earlier proposal of Heyting [50] to add a provability operator to the language of intuitionistic logic. van Atten [108] for additional details the development of Gödel's interpretation and its relationship to the BHK interpretation.

³See the introduction to [35] for a summary of other related results.

$$(2.5) \quad (\mathcal{B}\varphi)^* = \text{Prov}_T(\ulcorner \varphi^* \urcorner)$$

But note that

$$(2.6) \quad \mathbf{S4} \vdash \mathcal{B}(\mathcal{B}\varphi \rightarrow \varphi)$$

Taking $\varphi \equiv \perp$ we have that $\mathbf{S4} \vdash \mathcal{B}(\neg\mathcal{B}\perp)$ which, interpreted relative to (2.5), expresses that the consistency of T is provable within T itself. As Gödel pointed out, if T is consistent and sufficiently strong to derive the Second Incompleteness Theorem then $\mathcal{B}$ cannot be interpreted as expressing provability in T .⁴

This limitation notwithstanding, Gödel’s proposal that a modal operator satisfying the axioms of **S4** can be interpreted as expressing something like informal provability or demonstrability has been quite influential. For suppose that we follow the view about modal logic championed by Quine (e.g. [87]) and hold that the only coherent way of interpreting “It is necessary that” in natural language is as logical necessity (i.e. “what is self-contradictory to say is otherwise”) as opposed to metaphysical necessity (i.e. “what could not have been otherwise”). Burgess [13] has proposed that on this understanding

⁴This is an important historical observation because the system for which Montague and Kaplan originally framed the Knower Paradox includes an axiom (known as U) corresponding to the interpretation of (2.6) under $(\cdot)^*$. A variety of commentators starting with Anderson [1] have argued that the locus of the paradoxes lies with this principle and suggested reasons why it should be given up. However, it appears that Gödel realized the problems involved with providing an arithmetic interpretation of (2.6) in the early 1930s.

of modality the question “which is the right logic of necessity?” naturally splits into questions based on whether logical necessity is understood model theoretically (i.e. as validity or truth in all models) or proof theoretically (i.e. as demonstrability or the existence of a verifying derivation employing only logical principles). Understood in the latter sense, we are faced with the task of identifying which principles involving the $\mathcal{B}$ operator ought to be regarded as valid when the notion of demonstrability is not assumed to coincide with derivability in a particular formal system. Expanding on arguments of Haldén [45], Burgess argues that **S4** is the correct logic of informal demonstrability.⁵

But one might also think that however strong the case in favor of **S4** as the correct logic of informal provability, the treatment of this notion as a propositional operator is itself not completely satisfactory. The most obvious reason for this is that although it is clear that demonstrability *can* function as a sentential operator (as in “If φ is provable, then it is provable that φ ”

⁵Many of these arguments mirror those which can be given in favor of adopting the LP as principles about the explicit proof/theorem relation sketched in Chapter 1. For instance, in favor of the axiom $\mathcal{B}\varphi \rightarrow \varphi$ we may argue that if φ has been demonstrated by purely logical means – i.e. derived from purely logical axioms by logically sound rules of inference – then it certainly cannot be false. As Burgess points out, however, establishing that **S4** is *the* correct logic of demonstrability also requires that we provide an argument *against* the characteristic axioms of supersystems such as **S5** or **S4** together with McKinsey or Gregorczyk axiom when $\mathcal{B}$ is interpreted as demonstrability. Following Haldén, Burgess suggests that such arguments can be given on the basis of additional plausible principles about demonstrability – e.g. that we can demonstrate that there are statements of sentential logic which are neither tautologies nor contradictions.

is provable”), the notion informal probability itself still implicitly involves quantification over some notion of proof or demonstration. For to say that φ is *provable* (*demonstrable*) presumably just means that there exists a proof (demonstration) which establishes φ .⁶ This in turn is to say that any formal analysis we propose of (2.1) ought to at least be compatible with one on which $\mathcal{B}$ is analyzed according to

$$(2.7) \quad \mathcal{B}\varphi \quad \Rightarrow \text{there exists a proof of } \varphi$$

It is presumably implicit in our background understanding of provability that we ought to count a statement φ as provable not just in the case where we have already proven φ or even in the case we might might feasibly prove φ taking limitations of time or resources into account. Rather to say that φ is provable presumably means that φ can be demonstrated logically from principles which we accept, no matter how long or complex its derivation might be. From this it follows that the existential quantifier on the right-hand side of (2.7) ought to be understood as ranging over all those derivations

⁶This is a feature which distinguishes the two senses of logical necessity mentioned above. For note if we analyze “ φ is logically valid” in the familiar way as “ φ is valid in virtue of logical form” and then analyze the latter as “all sentences obtainable by replacing the non-logical constituents of φ are true” we are lead naturally to think of the former sort of statement as expressing implicit *universal* quantification over something like valuations or models. This feature of our concept of validity leads naturally to the standard relational (i.e. Kripke) semantics for the modal logic in which $\Box\varphi$ holds at a world just in case φ holds at *all* accessible worlds. This observation raises the question of what a conceptually motivated “existential” semantics for the demonstrability modality might look like.

which we might in principle construct and recognize as proofs.⁷

Gödel [37] seems to have been inspired by both of the foregoing considerations when he proposed that rather than attempting to axiomatize the properties satisfied by a provability modality, we attempt to provide an interpretation of intuitionism by axiomatizing the properties satisfied by the relation *z is a derivation of φ from ψ* . He states that “with good enough will” (p. 101) this relation – for which he adopts the notation $zB\varphi, \psi$ – can be regarded as decidable. Although Gödel does not present a fully worked out formal system for reasoning about this notion, he does propose the following axioms involving B :

$$(2.8) \quad \text{a) } (zB\varphi, \psi \wedge uB\psi, \chi) \rightarrow f(z, u)B\varphi, \psi$$

$$\text{b) } zB\varphi(x, y) \rightarrow \varphi(x, y)$$

$$\text{c) } uB\varphi \rightarrow u'B(uB\varphi)$$

$$\text{d) If } \varphi \text{ has been derived with proof } a, \text{ then infer } aB\varphi.$$

⁷Leitgeb [65] makes a similar argument for interpreting an informal provability modality as expressing “provability in principle” in a manner which thereby requires us to make sense of the notion of a proof which we could construct independent of resource limitations. As we will see below, however, there is an important sense in which our desire to impose such an interpretation on quantification over proofs does not by itself fix the intended range of quantifiers in formal systems which can be constructed to formalize the interpretation (2.7). In fact, I will argue below that the epistemic paradoxes can both be seen as highlighting the significance of this point.

In their introduction to [37], Sieg and Parsons make the following observations about these axioms: 1) the context suggests that the atomic formulas which may appear in the language under consideration are propositional compounds which may involve B and variables intended to denote proofs; 2) Gödel's use of B as a two place relation $zB\varphi$ can be understood to abbreviate $zB\top\varphi$ where $\top$ is some fixed tautology. Taking these observations into account and taking $\varphi \equiv \top$ in (2.8a) then we have

$$(2.9) \quad (zB\psi \wedge uB\psi, \chi) \rightarrow f(z, u)B\psi$$

This not only bears a strong resemblance to the application axiom of **LP**, but it can also be taken as an attempt to formalize the BHK understanding of the condition (2.2b) discussed above. Similar remarks apply to (2.8b) and the **LP** reflection axiom, (2.8c) and the **LP** proof checker axiom and (2.8d) and the Constructive Necessitation Rule to be discussed below.

Although he does not say so explicitly, it is also clear that Gödel intends the system to allow for quantification over proofs. For instance, he states that it is possible to derive the following statement:

$$(2.10) \quad aB[(\forall x)\neg xB(0 = 1)]$$

However, Gödel does not provide the derivation of this statement in the system. This leaves open two questions: 1) what principles about the interaction

of proof quantifiers and the B relation must be assumed in order to derive (2.10) in the sort of system in question; 2) is the derivability of this statement a desirable feature of the system.

In response to question 1), Sieg and Parsons note that it is clearly possible to derive $\neg xB(0 = 1)$ in Gödel's system. From this

$$(2.11) \quad a'B(\neg xB(0 = 1))$$

for some proof term a' by the rule (2.8d). But it is not clear how (2.10) is derivable from this by standard universal generalization. For not only does Gödel also indicate that formulas in the scope of B appear in “in quotes” (and thus not open to quantification) but also the proof term a in (2.10) presumably has to subsume the fact this formula has been derived from (2.11) by universal generalization. The former problem could presumably be solved by the use of a syntactic device akin to Feferman's $\ulcorner \cdot \urcorner$ notation by which quantifiers can bind variables appearing in the (names of) formulas inside a godel numbering function through the use of an arithmetically defined substitution operator. I will suggest below that the latter problem is at least clarified by the use of a system of explicit modal logic like Fitting's system QLP. However, the adoption of such a system also points to several respect in which the inference from (2.11) to (2.10) might be problematic.

Again, Gödel appears to foresee the difficulties which the derivability of $aB[(\forall x)\neg xB(0 = 1)]$ raises for what he refers to as the “modal logic” route to interpret constructive reasoning. Among these is whether the sorts of proofs to which B can be applied can be taken to be “surveyable” – i.e. inscribable within a constructively presented formal system such as the outlined system itself. Here Gödel appears to formulate a dilemma. On the one hand, if the question is answered positively (and we again assume that the system contains a sufficient amount of arithmetic – a condition which again is not clearly stated in Gödel’s exposition), then we should expect that (2.10) should *not* be derivable for essentially the same reasons given in regard to (2.6) above. On the other hand, if the question is answered in the negative, it is not at all clear whether the intended system should be taken to satisfy the minimal standards which Gödel argues are necessary features of an analysis of constructive reasoning.⁸

⁸Gödel’s treatment of intuitionism in [37] occurs inside a broader consideration of how constructive reasoning can be formalized in order to provide secure consistency proofs for arithmetic and analysis in light of the Second Incompleteness Theorem. One proposal he considered is to use proof theoretic techniques like the negative translation to embed classical systems (e.g. PA) into intuitionistic ones (e.g. HA) and then argue directly for the consistency of the intuitionistic systems. Gödel suggests that if we follow this route, we will face the additional problem of analyzing Heyting’s interpretation of the logical connectives. He proposes that this can be carried out either by an embedding into a propositional modal system (as in his [35] interpretation) or into a system of proto-explicit modal logic of the sort just described. The fact that we reach a problem of interpretation in both cases appears to provide the basis for his dismissive reaction to this approach (“this is the worst of the three ways . . .”). See Tait [101] for further appraisal of Gödel’s critique.

The foregoing remarks set the stage for the introduction of explicit modal logic of the sort previewed in Chapter 1. For on the one hand, we have seen that although it has been traditional to regard informal provability or demonstrability as a modality, there is also a clear motivation for regarding statements of the form φ *is provable* as expressing quantification over proofs even if the relevant notion of proof is still understood independently of any particular formal system. Once this is acknowledged, it becomes natural to attempt to formalize the intuitively valid properties satisfied by the relation “ t is a proof of φ .” And as Gödel appears to have observed, a schematic understanding of this sort is compatible with the introduction of quantifiers which range over proofs, while at the same leaving open the questions about the domain over which they can be taken to vary and the logical principles they can be taken to obey.

These observations prefigure important waypoints in the development and justification of Artemov’s Logic of Proofs and Fitting’s Quantified Logic of Proofs. In the subsequent sections of this chapter will I present the system LP as well as several variants of QLP. In each case, I will present the syntax, axiomatization and at least one form of semantics together with some basic results about the proof theory for the system in question. Most of the results cited here are not new. I have consequently elected to keep the exposition as

compact as possible by making reference to the relevant literature rather than reproducing proofs and auxiliary definitions which will not be required by my subsequent reconstruction of the epistemic paradoxes in Chapters 3 and 4. Two exceptions to the foregoing are a discussion in Section 3.2 of different ways in which reasoning by universal generalization can be internalized in a quantified extension of LP and a discussion in Section 3.5 of explicit analogs to the Barcan and Coverse Barcan-formulas.

2.2 The Logic of Proofs

The Logic of Proofs was presented by Artemov over a sequence of papers culminating in [3]. Many variants of the basic system presented there have now been presented and studied extensively – cf., e.g., [62] for a reasonably up to date survey.

2.2.1 The language of LP

As we have seen, the language of LP is based on a propositional language containing so-called *explicit modalities* of the form $t : \varphi$ with the intended interpretation

$$(2.12) \quad t : \varphi \quad \rightleftharpoons \quad t \text{ is a proof } \varphi$$

In such statements, t is a potentially structured expression known as a *proof term* drawn from an inductively defined class of terms defined as follows.

Definition 2.2.1. The class of LP proof terms is specified by the grammar

$$t := x_i \mid a_i \mid !t \mid t_1 \cdot t_2 \mid t_1 + t_2$$

$x_1, x_2, \dots$ are known as *proof variables*, $a_1, a_2, \dots$ as *axiom constants*. $!$, $\cdot$, $+$ and denote *proof operations* respectively called *proof checker* (unary), *application* (binary) and *sum* (binary).

The formulas of LP are now defined very much as formulas in propositional modal logic with the exception that that proof terms rather than traditional modal operators are now prefixed to formulas using the $:$ notation of (2.12).

Definition 2.2.2. If $P_0, P_1, \dots$ are propositional letters, then the class of formulas of LP is specified by the grammar

$$\varphi := \perp \mid P_i \mid \varphi_1 \wedge \varphi_2 \mid \varphi_1 \vee \varphi_2 \mid \varphi_1 \rightarrow \varphi_2 \mid \neg\varphi_1 \mid t : \varphi$$

As we have seen, the intended interpretation of an explicit modality $t : \varphi$ is “ t denotes a proof of φ .” Both the arithmetic and relational semantics presented below attempt to make this interpretation precise according to different background notions of proof.

2.2.2 Axiomatization

We may now specify a Hilbert-style axiom system for LP, first introducing the axioms and then the rules.

Definition 2.2.3. The axioms of LP are as follows:

- LP1 all tautologies of classical propositional logic
- LP2 $t : (\varphi \rightarrow \psi) \rightarrow (s : \varphi \rightarrow t \cdot s : \psi)$
- LP3 $t : \varphi \rightarrow \varphi$
- LP4 $t : \varphi \rightarrow !t : t : \varphi$
- LP5 $t : \varphi \rightarrow t + s : \varphi$ and $s : \varphi \rightarrow t + s : \varphi$

If Γ is a finite set of LP formulas, then the derivability relation $\Gamma \vdash_{\text{LP}} \varphi$ is specified by the following rules:

- RLP1 If $\varphi \in \Gamma$, then $\Gamma \vdash_{\text{LP}} \varphi$.
- RLP2 If $\Gamma \vdash_{\text{LP}} \varphi \rightarrow \psi$ and $\Gamma \vdash_{\text{LP}} \varphi$, then $\Gamma \vdash_{\text{LP}} \psi$.
- RLP3 If φ is an instance of LP1-LP5, then $\Gamma \vdash_{\text{LP}} \varphi$
- RLP4 If φ is an instance of one of LP1-LP5 and a an axiom constant, $\Gamma \vdash_{\text{LP}} a : \varphi$.

We write $\Gamma \vdash_{\text{LP}} \varphi$ if φ is derivable from premises Γ in this system, often omitting the subscript if context allows.

RLP4 is known as the *axiom necessitation rule* and is meant to reflect the idea that if φ is an instance of an LP axiom, then it is possible to specify a proof constant a which can serve as a name for a justification of φ during the course of a derivation. The application rule can be constrained by supplying a so-called *constant specification* $\mathcal{C}$ which assigns to each formula constant a one or more formulas $\mathcal{C}(a)$ which are taken to be the formulas for which a is taken to serve as justification. $\mathcal{C}$ is said to be *axiomatically appropriate* if

for each instance φ of an LP axiom, there is at least one constant a such that $\varphi \in \mathcal{C}(a)$, *schematic* if for each constant a , $\mathcal{C}(a)$ contains only instances of LP axiom schemes, *injective* if $\mathcal{C}(a)$ contains at most one formula for all a and *schematically injective* if $\mathcal{C}$ is schematic and each constant a , $\mathcal{C}(a)$ contains instances of a single axiom scheme. We say that a derivation of φ from hypotheses Γ is *based* on a constant specification $\mathcal{C}$ if all of its instances of $\vdash a : \psi$ of RLP3 are such that $\psi \in \mathcal{C}(a)$. Symbolically, this will be expressed as $\Gamma \vdash_{\text{LP}(\mathcal{C})} \varphi$ and the corresponding system denoted as $\text{LP}_{\mathcal{C}}$.

2.2.3 Basic results

LP enjoys the deduction theorem in the usual sense – i.e.

$$\Gamma, \varphi \vdash \psi \implies \Gamma \vdash \varphi \rightarrow \psi$$

Another interesting feature of this system is that it is able to internalize its own proofs in the following sense:

Proposition 2.2.1. If $\vdash_{\text{LP}} \varphi$, then $\vdash_{\text{LP}} t : \varphi$ for some proof term t .

This result is often referred as the *Constructive Necessitation Rule*. It can be taken as an explicit counterpart to the modal Necessitation Rule (i.e. $\vdash \varphi \therefore \vdash \Box\varphi$) which is common to all normal modal logics. Of course Proposition 2.2.1 is not a genuine rule as it does specify which proof term

t serves as a witness to the provability of φ . However, the standard proof of this result may be demonstrated in a manner which shows how t may be found effectively from a derivation of φ in **LP**. In fact, this is a special case of a more general result:

Theorem 2.2.1 (Lifting Lemma (Artemov [3])). Let $\vec{s} : \vec{\Gamma}$ denote $t_1 : \psi_1, \dots, t_n : \psi_n$, Δ denote $\chi_1, \dots, \chi_m$, and $\vec{y} : \vec{\Delta}$ denote $y_1 : \chi_1, \dots, y_m : \chi_m$ and suppose that $\vec{s} : \vec{\Gamma}, \Delta \vdash \varphi$. Then there exists a proof term $t(\vec{s}, \vec{y})$ such that

$$\vec{s} : \vec{\Gamma}, \vec{y} : \vec{\Delta} \vdash t(\vec{s}, \vec{y}) : \varphi$$

Moreover, if the original derivation is based on the an injective constant specification $\mathcal{C}$, then the resulting derivation of $t(\vec{s}, \vec{y}) : \varphi$ is based on an injective constant specification extending $\mathcal{C}$.

Proof. The proof proceeds by induction on derivations in **LP**. In the case where $\varphi \equiv s : \psi \in \vec{s} : \vec{\Gamma}$, let $t \equiv !s$. If $\varphi \equiv \chi_i \in \Delta$, then let $t = y_i$. If φ is an instance of an **LP** axiom, let $t = a$ for some new axiom constant not in $\mathcal{C}$ (in which case $\vdash a : \varphi$ by RLP4). If φ is derived by modus ponens from μ and ν , then by the induction hypothesis there are proof terms $u(\vec{s}, \vec{y})$ and $v(\vec{s}, \vec{y})$ such that $\vec{s} : \vec{\Gamma}, \vec{y} : \vec{\Delta} \vdash u(\vec{s}, \vec{y}) : \mu, v(\vec{s}, \vec{y}) : \nu$. Since it thus follows that $\vec{s} : \vec{\Gamma}, \vec{y} : \vec{\Delta} \vdash u \cdot v$ by LP2 and modus ponens, we may put $t \equiv u \cdot v$. If

φ is derived by an instance of RLP4, then $\varphi \equiv a : \psi$ for some LP axiom ψ .

As $a : \varphi \rightarrow !a : a : \varphi$ is also derivable by LP4, we may put $t = !a$. $\square$

As mentioned above, the intended interpretation of LP formulas of the form $t : \varphi$ is that the proof term denotes a proof or other form of justification of the formula φ . This ideology can be formalized in at least three different ways. The first of these builds a connection to Gödel's [35] embedding of IPC into S4 by showing how S4 derivations can be viewed as templates for LP derivations in which instances of the $\mathcal{B}$ modality are replaced with proof terms. The existence of such a connection should not be surprising given the similarity in form between LP2, LP3 and LP4 and the axioms K, T and 4 of S4. One half of this connection is expressed by the following proposition:

Proposition 2.2.2. Let $(\cdot)^0$ be the mapping which replaces all occurrences of $t : \psi$ occurring in formula φ with $\Box\psi$. Then if $\vdash_{\text{LP}} \varphi$, then $\vdash_{\text{S4}} \varphi^0$.

It is also possible to demonstrate a form of converse to Proposition 2.2.2 which provides a deeper connection between the Gödel interpretation of $\Box$ and the intended interpretation of explicit modalities. To this end, define a *realization* r of a propositional modal formula φ to be an assignment of proof terms to all occurrences of $\mathcal{B}$ in φ along with a constant specification for all constants appearing in these terms. We write φ^r for the image of φ under the

realization r . A realization is said to be *normal* if all negative occurrences of $\mathcal{B}$ in φ are replaced with proof variables and the corresponding constant specification is injective.

Theorem 2.2.2 (Realization Theorem (Artemov [3])). If $\vdash_{\mathbf{S4}} \varphi$, then $\vdash_{\mathbf{LP}} \varphi^r$ for some normal realization r .

If $(\cdot)^g$ is the Gödel embedding of IPC into §4, then it immediately follows that

$$\vdash_{\mathbf{IPC}} \varphi \iff \vdash_{\mathbf{LP}} (\varphi^g)^r \text{ for some normal realization } r$$

The Realization Theorem can thus be seen as providing a proof-based semantics for IPC which provides at least a partial analysis of the BHK interpretation of the intuitionistic connectives discussed above. For note that if we combine the foregoing result with the Gödel embedding of IPC into **S4** we obtain the following:

Theorem 2.2.3. $\mathbf{IPC} \vdash \varphi$ if and only if $\mathbf{LP} \vdash (\varphi^\circ)^r$ for some normal realization r .

If we take into account the fact that normal realization replaces instances of $\Box$ in negative positions with proof variables, then such a realization of a formula can be seen as making explicit functional dependencies between the proof of a complex statement and the proof of its constituents. For instance,

the LP application axiom LP2 is a realization of the S4 normality axiom $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$. LP2 can thus be viewed as formalizing the BHK interpretation of the conditional given by (2.2b) in the sense that it can be read as stating that if x is a proof of $(\varphi \rightarrow \psi)$, then x may be combined with any proof y of φ to yield a proof of ψ .

Decidability

The decidability of the validity problem $\text{LP}_{\mathcal{C}}$ can be shown to depend quite sensitively on the properties of the constant specification $\mathcal{C}$.

- 1) If the domain of $\mathcal{C}$ is finite, then $\text{LP}_{\mathcal{C}}$ is decidable [2].
- 2) If $\mathcal{C}$ is schematic, then $\text{LP}_{\mathcal{C}}$ is decidable [74].
- 3) There exists a decidable constant specification $\mathcal{C}$ for which $\text{LP}_{\mathcal{C}\mathcal{S}}$ is undecidable [61].

For a recent survey of decidability results about LP and related systems, see Kuznets [62].

2.2.4 Semantics

There are two other means of connecting LP with the intended interpretation (2.12) which take the form of semantic interpretations in a more

traditional sense. The first kind of semantics for **LP** was provided by Artemov [3] in the form of an interpretation $(\cdot)^*$ of $\mathcal{L}_{\mathbf{LP}}$ into the language $\mathcal{L}_a$ of first-order arithmetic described in Chapter 1. The other is a form of relational semantics due to Fitting [31] which combines the standard possible world models for **S4** with a more syntactic mechanism for keeping track of which formulas are justified by which proof terms at a given world. I will respectively refer to these as the *arithmetic* and *relational* semantics for **LP**. A presentation of the relational semantics will be postponed until Section 3 as it generalizes to **QLP** in such a way that its application to **LP** can be seen as a special case. However, the situation is somewhat more complex with respect to the arithmetic semantics. And I will hence present Artemov's original arithmetic semantics for **LP** here.

We will fix **PA** as an arithmetic background arithmetic theory and fix a canonical godel numbering $\ulcorner \cdot \urcorner : \text{Form}_{\mathcal{L}_A} \rightarrow \mathbb{N}$ which is injective. We also fix a multi-conclusion proof predicate $\text{Prf}(x, y)$ which is provably Δ_1 in **PA** and such that for every arithmetic formula φ

$$(2.13) \quad \mathbf{PA} \vdash \varphi \text{ if and only if for some } n \in \mathbb{N}, N \models \text{Prf}(\bar{n}, \overline{\ulcorner \varphi \urcorner}).$$

As described in [3] we will also assume that $T(n) = \{k \mid N \models \text{Prf}(\bar{n}, \bar{k})\}$ is finite for all n and that for all n, m there is a k such that $T(n) \cup T(m) \subseteq T(k)$.

The latter condition codifies the fact that the $Prf(x, y)$ is a multi-conclusion proof predicate in the sense that for any (codes of) proofs n, m which prove the set of conclusions $T(n)$ and $T(m)$, there is a (code of a) proof k which proves $T(n)$ and $T(m)$ simultaneously.

We must next provide a means of interpreting the proof operations $\cdot, +$ and $!$ of $\mathcal{L}_{\mathbf{QLP-}}$. This can be accomplished by defining recursive functions $m(x, y)$, $p(x, y)$ and $c(x)$ which compute operations on (codes of) arithmetic proofs corresponding to the intended interpretations of the $\mathbf{LP}$ proof operations. In particular, $m(x, y)$ returns the least godel-number of a proof containing all instances of ψ such that $\ulcorner \varphi \rightarrow \psi \urcorner \in T(x)$ and $\ulcorner \varphi \urcorner \in T(y)$. $p(x, y)$ returns the least godel-number of a proof containing $T(x)$ and $T(y)$ and $c(x)$ returns the least godel-number of a proof y such that $\text{Prf}(y, \overline{\ulcorner \varphi \urcorner}) \in T(y)$ for all $\ulcorner \varphi \urcorner \in T(x)$.

Definition 2.2.4. An arithmetic interpretation of $\mathbf{LP}$ is a function $(\cdot)^*$ which when applied to a proof term $t \in \text{Term}_{\mathbf{LP}}$ yields a natural number and when applied to a formula $\varphi \in \text{Form}_{\mathbf{LP}}$ yields a formula in the language of arithmetic. The value of $(\cdot)^*$ is defined inductively as follows:

- i) If $t \equiv x$ (i.e. a proof variable), then $(t)^* \in \mathbb{N}$
- ii) If $t \equiv t_1 \cdot t_2$, then $(t)^* = m(t_1^*, t_2^*)$

- iii) If $t \equiv !t_1$, then $(t)^* = c(t_1^*)$
- iv) If $t \equiv t_1 + t_2$, then $(t)^* = a(t_1^*, t_2^*)$
- v) $(\cdot)^*$ commutes with boolean connectives.
- vi) $(t : \varphi)^* = \text{Prf}(\overline{t^*}, \overline{\varphi^{*\top}})$

Note that under a given interpretation $(\cdot)^*$ all proof terms are mapped to natural numbers and all LP formulas are mapped to *closed* arithmetic formulas (this follows since proof variables are mapped to natural numbers). Note also that an arithmetic formula of the form $(t : \varphi)^*$ will always be provably Δ_1 .

Given a constant specification $\mathcal{C}$ a given $(\cdot)^*$ is said to be a $\mathcal{C}$ -*interpretation* if all φ such that $\varphi \in \mathcal{C}(a)$ for some axiom constant a are true. An LP formula is said to be *arithmetically valid* if $(\varphi)^*$ is true for all $(\cdot)^*$. φ is said to be *provably valid* (in PA) if $\text{PA} \vdash (\varphi)^*$ for any $(\cdot)^*$. We define analogs of these notions which can also be relativized to a constant specification: φ is said to be *arithmetically valid with respect to $\mathcal{C}$* if $(\varphi)^*$ is true for all $\mathcal{C}$ interpretations; φ is said to be *provably valid with respect to $\mathcal{C}$* if $(\varphi)^*$ is $\text{PA} \vdash (\varphi)^*$ for all $\mathcal{C}$ interpretations. Note that both the standard and $\mathcal{C}$ -relativized notions of provable validity imply the corresponding form of arithmetic validity.

Theorem 2.2.4. Arithmetic Soundness[Artemov [3]]

- 1) If $\text{LP} \vdash \varphi$, then φ^* is provably valid.
- 2) If $\text{LP}_c \vdash \varphi$, then φ^* is provably valid with respect to $\mathcal{C}$.

Proof. Since 2) obviously implies 1), we reason by induction on derivation in LP_c . We must first show that each for each axiom φ of LP , $\text{PA} \vdash (\varphi)^*$. An illustrative case is $\varphi \equiv t : \psi \rightarrow \psi$. There are two cases to consider: either i) $N \models (t : \varphi)^*$ or ii) $N \not\models (t : \varphi)^*$. In the first case, $(t : \varphi)^* = \text{Prf}(t, \ulcorner \varphi^* \urcorner)$ is a true Δ_1 formula from which it follows that $\text{PA} \vdash \varphi^*$ and hence also $\text{PA} \vdash \text{Prf}(t, \ulcorner \varphi^* \urcorner) \rightarrow \varphi^* \equiv (t : \varphi \rightarrow \varphi)^*$. In the second case, $(t : \varphi)^*$ is a false Δ_1 formula and hence $\text{PA} \vdash \neg(t : \varphi)^*$. But then $\text{PA} \vdash (t : \varphi \rightarrow \varphi)^*$. Since the provability of the image under $(\cdot)^*$ of all LP axioms may be established in this way, it follows that the axiom necessitation rule preserves provability in PA . $\square$

It is also possible to establish the converse – i.e.

Theorem 2.2.5. Arithmetic Completeness[Artemov [3]] If $\vdash_{\text{LP}(c)} \varphi$, then φ is $\mathcal{C}$ provably valid with respect to $\mathcal{C}$

2.3 The Quantified Logic of Proofs

A quantified form of the Logic of Proofs was first studied by Yavorsky [118]. Several notions of arithmetic validity were defined and it was shown that the resulting sets of valid formulas were highly undecidable.⁹ Fitting independently introduced two axiomatizations of a quantified logic of proofs in [29] and [32] along with two forms of relational semantics which generalize his previous [31] semantic treatment of LP. These systems differ in the way in which they handle internalization the of inferences mediated by the application of the Universal Generalization rule when applied to a formula containing a free proof variable. The distinction between them is somewhat subtle and as we will see below, it is not clear whether our intuitions about explicit modalities and quantification determinately decide between them. The axiom system which I will hereafter refer to as QLP is a minor modification of that presented in [32] in that it employs a more general form of what he refers to as the Justified Universal Generalization rule. It will be shown below that this system has the same set of theorems as that

⁹It is difficult to provide a direct comparison of the system presented by Yavorsky with the forms of QLP which will be described here because the language he employs does not contain proof operations. The systems themselves are also not presented axiomatically (out of necessity since they are at minimum Π_2 hard), but rather in terms of the class of sentences which are provable under all arithmetic interpretations in some theory T extending PA.

presented in [32].

2.3.1 The language of QLP

The language $\mathcal{L}_{\text{QLP}}$ of QLP extends the language of LP in three ways:

- i) Rather than containing axiom constants, $\mathcal{L}_{\text{QLP}}$ possesses so-called *primitive proof terms* of the form $a(x_1, \dots, x_n)$ where a is a primitive functional symbol and $x_1, \dots, x_n$ are free proof variables.
- ii) $\mathcal{L}_{\text{QLP}}$ contains an additional binary proof operation $(\cdot \forall \cdot)$ known as the *uniform verifier*. The operation takes a proof term t as its first argument and a proof variable x as its second argument.
- iii) The language of QLP extends that of LP by allowing quantification over proof variables so that if $\varphi(x)$ is a formula, then so are $(\forall x)\varphi(x)$ and $(\exists x)\varphi(x)$.

The free variables of a proof term t are defined in the same manner as those of functional expressions in first-order logic with the exception that x is considered bound in $(t\forall x)$. Precisely: $\text{FV}(x_i) = \{x_i\}$, $\text{FV}(a(x_1, \dots, x_n)) = \{x_1, \dots, x_n\}$, $\text{FV}(t \cdot s) = \text{FV}(t + s) = \text{FV}(t) \cup \text{FV}(s)$, $\text{FV}(!t) = \text{FV}(t)$, $\text{FV}((t\forall x)) = \text{FV}(t) \setminus \{x\}$. The free variables of a QLP formula φ are defined exactly as in first-order logic – i.e. $\text{FV}(P_i) = \emptyset$, $\text{FV}(t : \psi) = \text{FV}(t) \cup \text{FV}(\psi)$,

$FV(Qx\psi) = FV(\psi) \setminus \{x\}$ where Q is $\forall$ or $\exists$.

2.3.2 Axiomatization

Over the course of this subsection I will introduce four systems for formalizing first-order reasoning about proofs: QLP_0 , QLP_1 , QLP_2 and QLP . The first of these will be a system which lacks a mechanism for internalizing universal generalization (and for which the corresponding analog of the Constructive Necessitation Theorem fails). The third and second of these will respectively correspond to the system of [32] and a minor variant of the system of [31] (wherein proof constants are replaced with primitive proof terms and definitions involving the constant specification are accordingly replaced with ones involving primitive term specifications). The fourth will be my proposed reformulation of the system of [32].

Basic quantifier axioms and rules

Definition 2.3.1. The axioms of QLP_0 include LP1-LP5 together with

QLP1 $(\forall x)\varphi(x) \rightarrow \varphi(t)$, for any proof term t that is free for x in $\varphi(x)$

QLP2 $\varphi(t) \rightarrow (\exists x)\varphi(x)$, for any proof term t that is free for x in $\varphi(x)$

QLP3 $(\forall x)(\psi \rightarrow \varphi(x)) \rightarrow (\psi \rightarrow (\forall x)\varphi(x))$, where $x \notin FV(\psi)$

QLP4 $(\forall x)(\varphi(x) \rightarrow \psi) \rightarrow ((\exists x)\varphi(x) \rightarrow \psi)$, where $x \notin FV(\psi)$

It will be recognized that these axioms correspond precisely in form to the standard Hilbert axioms for first-order logic (cf., e.g., [104]).

In order to state the rule form of $\mathbf{QLP}_0$, we must first provide several auxiliary definitions which generalize the notion of a constant specification for LP. The general idea is that since free variables will be open to quantification in QLP, the proof constant associated with an axiom ought to reflect which free variables it contains. Following Fitting [32], I will adopt the expression *primitive term specification* to refer to the generalized notion of constant specification. More precisely, such a primitive term specification is a mapping $\mathcal{P}$ which assigns to each primitive proof term a a (possibly empty) set of formulas $\mathcal{P}(a)$ such that if $\varphi \in \mathcal{P}(a)$, then $\text{FV}(a) = \text{FV}(\varphi)$. Such a specification is said to *meet the free variable condition* if whenever $\varphi(x_1, \dots, x_n) \in \mathcal{P}(a(x_1, \dots, x_n))$ and $y_1, \dots, y_n$ are variables which do not occur in $\varphi(x_1, \dots, x_n)$, then $\varphi(y_1, \dots, y_n) \in \mathcal{P}(a(y_1, \dots, y_n))$. $\mathcal{P}$ is *axiomatically appropriate* if for all (and only) instances φ of the QLP axioms, there exists a primitive proof term a such that $\varphi \in \mathcal{P}(a)$.

Definition 2.3.2. The rules of $\mathbf{QLP}_0$ include RLP1-2 together with

RQLP1 If φ is an instance of LP1-LP5 or QLP1-QLP4 then $\Gamma \vdash_{\mathbf{QLP}_0} \varphi$.

RQLP2 If $\Gamma \vdash \varphi(x)$, then $\Gamma \vdash (\forall x)\varphi(x)$ if $x \notin \text{FV}(\Gamma)$.

RQLP3 If $\varphi \in \mathcal{P}(a)$, then $\vdash a : \varphi$.

Rules RQLP1 and RQLP2 will subsequently be referred to as *Universal Generalization* [UG] and *Axiom Necessitation*. With these rules in place, we can

now define the system $\mathbf{QLP}_0$ to be the system consisting of axioms LP1-LP5, QLP1-QLP4 and rules RQLP1-RQLP4. Following the notation of Section 2, I will write $\Gamma \vdash_{\mathbf{QLP}_0} \varphi$ to denote that φ is derivable in $\mathbf{QLP}_0$ and $\Gamma \vdash_{\mathbf{QLP}_0(\mathcal{P})} \varphi$ to denote that φ is derivable in $\mathbf{QLP}_0$ where all instances of Axiom Necessitation take the form $a : \varphi$ where a is the domain of $\mathcal{P}$. $\mathbf{QLP}_0(\mathcal{P})$ will be used to denote the corresponding deductive system.

Forms of Justified Universal Generalization

The Lifting Lemma demonstrates the sense in which LP is able to internalize reasoning about its own proofs. This feature is a central characteristic of explicit modal logics and it is thus desirable that an analogous result should hold for any reasonable extension of LP with quantifiers. Recall that the proof of the Lifting Lemma proceeds by induction on LP derivations, with a case for each LP rule. The only essentially new case is that of the $\mathbf{QLP}_0$ Universal Generalization rule. The question which we must address is how to extend $\mathbf{QLP}_0$ so as enable the internalization of steps mediated by the application of this rule.

To begin, note that the UG rule of QLP matches that form of the standard universal generalization rule of first-order logic. The justification which is standardly provided for this is as follows: if we are to derive $\varphi(x)$ from a

set of premises Γ which does not contain any assumptions about the object denoted by x , then it must be the case that φ holds for all objects. Since it is part of the ideology of explicit modal logic to regard proofs as objects in their own right, we ought to expect that this mode of reasoning ought to remain valid. It is, for instance, exactly such a rule which is required to derive the generalized forms of the LP axioms mentioned in Chapter 1.

By way of example, consider the following derivation in QLP_0 of a prenex normal form of the proof checker axiom:

- i) $x : \varphi \rightarrow !x : x : \varphi$ LP4
- ii) $!x : x : \varphi \rightarrow (\exists y)y : x : \varphi$ QLP2
- iii) $x : \varphi \rightarrow (\exists y)y : x : \varphi$ i), ii)
- iv) $(\exists y)[x : \varphi \rightarrow y : x : \varphi]$ QLP4, propositional logic from iii)
- v) $(\forall x)(\exists y)[x : \varphi \rightarrow y : x : \varphi]$ UG, iv)

Suppose we now wish to obtain a proof term t for which we can show

$$(2.14) \quad t : (\forall x)(\exists y)[x : \varphi \rightarrow y : x : \varphi]$$

Using the Lifting Lemma, we can obtain a term s such that we can prove an internalized version of iv) – i.e.

$$(2.15) \quad s(x) : (\exists y)[x : \varphi \rightarrow y : x : \varphi]$$

The question is what sort of rule we should adopt so as to enable us to find a term t by the inference from step iv) to step v) via UG can be internalized.

One option is to first derive

$$(2.16) \quad (\forall x)s(x) : (\exists y)[x : \varphi \rightarrow y : x : \varphi]$$

from (2.15) via UG and then to adopt an axiom which allows us to transpose quantifiers and explicit modalities. Such an axiom could take the general form

$$(UBF) \quad (\forall x)t(x) : \varphi(x) \rightarrow \langle t\forall x \rangle : (\forall x)\varphi(x)$$

where $\langle t\forall x \rangle$ is intended to denote a new proof operation which in this instance takes as arguments the proof term t which justifies the open formula $\varphi(x)$ and a variable x relative to which the generalization is performed. In order to justify the adoption of such an axiom, it is useful to contrast it with the formula

$$(EBF) \quad (\forall x)(\exists y)y : \varphi(x) \rightarrow (\exists y)y : (\forall x)\varphi \text{ (where } y \text{ does not occur free in } x)$$

which states that if for all justifications x , there is a justification y which justifies $\varphi(x)$, then there is a single justification denoted by y which serves to justify the universal generalization of $\varphi(x)$.

This latter principle seems implausible regardless of whether we think of QLP_0 quantifiers as ranging over formal or informal proofs. For even if we knew that for each proof x there was a corresponding proof y which served as evidence for $\varphi(x)$, this would not guarantee that we possess a means of subsuming these individual proofs under a single proof which serves to justify $(\forall x)\varphi(x)$.¹⁰ UBF differs from EBF precisely in that its antecedent requires that for all x , the proof denoted by $t(x)$ serves to justify $\varphi(x)$, thereby providing a uniform method for justifying $\varphi(x)$ for an arbitrary proof x . On the basis of this observation, Fitting adopted UBF into his original formulation of QLP in [31], referring to it in virtue of its form as the *Uniform Barcan Formula*. It may readily be verified that the system $\text{QLP}_1 = \text{QLP}_0 + \text{UBF}$ satisfies an analog of the Lifting Lemma wherein UBF is used to internalize steps mediated by UG in the manner suggested above.

If this were the only feature constraining the development of a quantified extension of LP, we might reasonably stop here. There are, however, reasons to think that UBF is not an appropriate principle by which to achieve this end. For note on the one hand that the premise to the UG rule requires not just that $\varphi(x)$ is true, but that it is provable within the system in question.

¹⁰In fact, as Fitting [32] notes, such a principle would amount to an informal form of the arithmetic ω -rule.

One might expect that the same significance was conveyed by $t(x) : \varphi(x)$ in the antecedent of UBF. However, since UBF is as an axiom rather than a rule, it has the effect of allowing us to infer $\langle t(x)\forall x \rangle : (\forall x)\varphi(x)$ in any situation where $(\forall x)t(x) : \varphi(x)$ is also true, even if it is not provable. To see that truth and provability in an appropriate extension of QLP_0 need not coincide, we need to informally consider the relational semantics I will present in the next section.

Relative to these semantics, a formula of the form $(\forall x) : \varphi(x)$ is taken to be true at a world w just in case $\varphi(x)$ holds for each object (i.e. justification) in the domain $\mathcal{D}(w)$ of w . However, an explicit modality of the form $t : \varphi$ is taken to hold at w just in case φ is true at all worlds accessible from w and t is taken to count as evidence for φ at w (the latter condition is handled in a syntactic manner which will be explained below). We might thus imagine a situation in which $t(x) : \varphi(x)$ holds for all $r \in \mathcal{D}(w)$ (which in turn means that $\varphi(x)[x/r]$ must be true at all accessible worlds), but in which $(\forall x)\varphi(x)$ still fails at an accessible world w' in virtue of the existence at w' of a justification r' not in the domain of w for which $\varphi(x)$ fails. However, in such a case UBF would still imply that $\langle t\forall x \rangle : (\forall x)\varphi(x)$ was true at w which would in turn wrongly imply that $(\forall x)\varphi(x)$ was true at w' . In the context where the class of available justifications is allowed to vary from

world to world, the antecedent of UBF thus appears to express only the “local” verifiability of $\varphi(x)$ and is hence too weak to allow us to derive a “global” statement with the form of its consequent.

In order to resolve this problem, Fitting [32] proposed a new version of QLP in which UBF was dropped in favor of another axiom and an additional rule which he respectively refers to as the *Uniformity Formula* and the *Justified Universal Generalization Rule*:

$$(UF) \quad (\exists y)y : (\forall x)t(x) : \varphi(x) \rightarrow \langle t(x)\forall x \rangle : (\forall x)\varphi(x)$$

$$(JUG) \quad \vdash t(x) : \varphi(x) \quad \therefore \quad \vdash \langle t(x)\forall x \rangle : (\forall x)\varphi(x)$$

I will refer to the system consisting of QLP_0 together UF and JUG as QLP_2 . JUG might at first appear to be a reformulation of UBF as a rule. However, in this case the requirement is that the antecedent $t(x) : \varphi(x)$ be provable in the system itself rather than merely true. But it turns out that if we adopt only JUG and not UF, the system is not complete with respect to the relational semantics alluded to above. As Fitting [32] demonstrates, the gap can be closed by taking UF as an axiom. However, the motivation for accepting this principle as an axiom is somewhat obscure once we have already adopted JUG.¹¹

¹¹Note that UF differs from UBF in the presence of the outside proof quantifier $(\exists y)y :$

The resolution I propose is to adopt the following generalized form of JUG which allows for necessitated hypotheses in which x does not appear free¹²:

$$\begin{aligned} (\text{GJUG}) \quad & \vec{s} : \vec{\Gamma} \vdash t(x) : \varphi(x) \quad \therefore \quad \vec{s} : \vec{\Gamma} \vdash \langle t(x) \forall x \rangle : (\forall x) \varphi(x) \text{ where} \\ & x \notin \text{FV}(\vec{s} : \vec{\psi}) \end{aligned}$$

This rule may be motivated by a comparison with the forms of the Universal Generalization rule for first-order order logic and the Necessitation rule for S4 which allow hypotheses – i.e.

$$(\text{foUG}) \quad \Gamma \vdash \varphi(x) \quad \therefore \quad \Gamma \vdash (\forall x) \varphi(x) \quad \text{if } x \notin \text{FV}(\Gamma)$$

$$\begin{aligned} (\text{SNec}) \quad & \Box \Gamma \vdash \varphi \quad \therefore \quad \Box \Gamma \vdash \Box \varphi \quad \text{where all } \psi \in \Box \Gamma \text{ are of the form} \\ & \psi \equiv \Box \chi \end{aligned}$$

The condition on Γ in foUG can be viewed as formalizing the fact that we are justified in concluding $(\forall x) \varphi(x)$ from a proof of $\varphi(x)$ as long as this proof has not relied on any premises about the object denoted by x . Similarly, the condition on Γ in **SNec** can be viewed as formalizing the fact that

in its antecedent. The antecedent of UF can thus be taken as expressing the provable the provability of $(\forall x) t(x) : \varphi(x)$ in the language of QLP. However, within the relational semantics which will be presented below, a gap between the truth of such a statement and its axiomatic derivability may still exist. This is also evident from the fact that UF is not already provable from JUG alone.

¹²This rule was independently proposed by Hidenori Kurokawa.

we are justified in concluding that φ is necessary from a proof of φ if this proof has not relied only on premises which are not themselves necessary. Combining these two principles together, it should follow that if we are able to derive a statement of the form $t(x) : \varphi(x)$ expressing that $t(x)$ denotes a justification of $\varphi(x)$ in a manner which does not rely on any premises about the justification denoted by x or which are not themselves justified, then this proof – as encoded in the structure of $t(x)$ – ought to serve as a justification of $(\forall x)\varphi(x)$.

I will refer to the system QLP_0 together with GJUG as QLP . It, together with QLP_0 , will serve as the basic system in which I will go about reconstructing the epistemic paradoxes in the following chapters. With this said, it should be pointed out immediately that the difference between QLP and QLP_2 lies primarily at the level of justifying their axioms and rules as it may be shown that they have the same class of theorems. The inclusion of QLP in QLP_2 is most readily demonstrated using the relational semantics for QLP which will be presented below. However, the other direction of this claim follows easily upon noting that GJUG is a special case of JUG together with the following:

Proposition 2.3.1. QLP derives all instances of UF.

Proof. We may reason in **QLP** as follows:

- i) $y : (\forall x)t(x) : \varphi(x) \vdash (\forall x)t(x) : \varphi(x)$ RLP1, LP3
- ii) $y : (\forall x)t(x) : \varphi(x) \vdash (\forall x)t(x) : \varphi(x) \rightarrow t(x) : \varphi(x)$ QLP1
- iii) $y : (\forall x)t(x) : \varphi(x) \vdash t(x) : \varphi(x)$ i), ii)
- iv) $y : (\forall x)t(x) : \varphi(x) \vdash \langle t\forall x \rangle : (\forall x)\varphi(x)$ GJUG
- v) $\vdash y : (\forall x)t(x) : \varphi(x) \rightarrow \langle t\forall x \rangle : (\forall x)\varphi(x)$
- vi) $\vdash (\exists y)y : (\forall x)t(x) : \varphi(x) \rightarrow \langle t\forall x \rangle : (\forall x)\varphi(x)$ UG, QLP4

□

Finally, we should record the fact that the adoption of GJUG is sufficient to prove internalization for **QLP**:

Proposition 2.3.2 (Lifting Lemma for **QLP**). If $\vec{s} : \vec{\Gamma}, \Delta \vdash \varphi$. Then there exists a proof term $t(\vec{s}, \vec{y})$ such that

$$\vec{s} : \vec{\Gamma}, \vec{y} : \vec{\Delta} \vdash_{\mathbf{QLP}} t(\vec{s}, \vec{y}) : \varphi$$

Proof. We extend the inductive proof given for **LP** by considering the additional case where $\vec{s}, \Delta \vdash (\forall x)\varphi(x)$ is derived by UG in **QLP**. In this case we have that x is not among the free variables of $\vec{s} : \vec{\Gamma}$ or Δ and $\vec{s}, \Delta \vdash \varphi(x)$. But

then $\vec{s} : \vec{\Gamma}, \vec{y} : \Delta \vdash t(\vec{s}, \vec{y}) : \varphi(x)$ by the induction hypothesis and therefore $\vec{s} : \vec{\Gamma}, \vec{y} : \Delta \vdash \langle t\forall x \rangle : (\forall x)\varphi(x)$ follows by GJUG. $\square$

2.3.3 Semantics

Unlike LP, QLP was first investigated using a relational semantics rather than an arithmetic one. This treatment extends Fitting's [29] earlier relational semantics for LP. The characteristic feature of this semantics is that it places two requirements on what must be true of an explicit modality $t : \varphi$ to be true at a world w : 1) φ must be true at all accessible worlds; 2) t must be acceptable evidence for φ at w . The first of these requirements is identical to the truth conditions for a formula of the form $\Box\varphi$ in the standard relational semantics for traditional modal logic. As such, condition 1) can be cashed out formally by using a standard relational structure for **S4** – i.e. a transitive and reflexive frame $\langle W, R \rangle$ and a propositional valuation V . If the relation R is taken to represent epistemic indiscernibility between worlds, then condition 1) can be taken to formalize the requirement that t should not be taken for evidence for φ at w for agent i if it is compatible with i 's knowledge that φ is false.

The second requirement on the truth of $t : \varphi$ at w is cashed out using a device known as an *evidence function* first employed by Mkrtychev [74].

In the case of LP, an evidence $\mathcal{E}$ is a binary mapping which assigns to each world w and proof term t a set of formulas for which t serves as evidence at w . In order to be an evidence function for a frame $\langle W, R \rangle$ such a mapping must satisfy a number of other structural conditions – e.g. the fact that a term counts as evidence for a formula is stipulated to be monotonic in the sense that if $R(w_1, w_2)$ and $\varphi \in \mathcal{E}(w_1, t)$, then $\varphi \in \mathcal{E}(w_2, t)$. However, no constraints are placed on whether the formulas for which an evidence function provides a justification at a world w must be *true* at w – i.e. it is compatible with semantics that $\varphi \in \mathcal{E}(w, t)$ at yet φ still be false at t . As such, Fitting [31] describes as $\mathcal{E}$ as providing *possible* evidence in the sense that if $\varphi \in \mathcal{E}(w, t)$, then t is potentially relevant to assessing the truth of φ (although possibly not decisive).

Although the semantics of [31] is not applied to a quantified system, it shares several features with the standard relational semantics for first-order modal logic (cf., e.g., [33], [54]). In particular, a relational model for LP can be thought of as possessing a domain in which LP proof terms take their values in each world and in which LP proof operations are assigned extensions. These features do not have to be made explicit in the semantics since it is possible to prove completeness using what are essentially term models – i.e. models in which the interpretation of an LP term is itself

and in which all worlds have the same domain. However, once we wish to treat proof variables as open to quantification, we are free to consider the case in which these restrictions are dropped – i.e. we may allow that the interpretation of a proof term in a model may denote an object other itself (and hence also that not all objects in the range of a proof quantifier must be named by proof terms) and that the class of such objects available may vary from world to world. Each of these possibilities is allowed relative to the semantics given by Fitting in [32], resulting in a very flexible notion of a QLP model which will prove useful in my subsequent treatment of the epistemic paradoxes. As we see below, however, this flexibility also means that we must face many of the same challenges in interpreting QLP formulas which arise in the interpretation of first-order modal logic.

Interpretations, evidence functions and models

As we will see below, the version of QLP presented here is sound and complete with respect to the same semantics given in [32]. The definition of QLP model which will be presented in this section thus corresponds exactly to that presented by Fitting [32] for QLP_2 to which the reader is referred for additional details and motivation.

A QLP model $\mathcal{M}$ is a structure consisting of the following six components:

- 1) a class of worlds W ;
- 2) a reflexive and transitive relation R on $W \times W$;
- 3) a *domain function* $\mathcal{D}$ which for every $w \in W$ assigns a non-empty set as the *domain* of w ;
- 4) an *interpretation function* $\mathcal{I}$ which assigns to each QLP proof operation a function of the appropriate type;
- 5) an *evidence function* $\mathcal{E}$ whose type and properties will be explained below;
- 6) a propositional *valuation* V mapping propositional letters into subsets of W .

A structure of the form $\langle W, R \rangle$ will be known as a *frame* and a structure of the form $\langle W, R, \mathcal{D} \rangle$ as an *augmented frame* in the case where $\mathcal{D}$ is monotonic with respect to R – i.e. if $R(w_1, w_2)$, then $\mathcal{D}(w_1) \subseteq \mathcal{D}(w_2)$. The class of $\overline{\mathcal{D}} = \bigcup_{w \in W} \mathcal{D}(w)$ will be referred to as the domain of the augmented frame.

As mentioned above, I will refer to its members – which will serve as the denotation of proof terms – as *reasons*.¹³

¹³QLP also possesses term models whose domain consists of syntactic entities much like those of a Herbrand model for first-order logic. However, nothing about the formulation of the relational semantics force such an interpretation. The term *reason* is thus a fitting description for a member of the domain of an arbitrary QLP model as it does not prejudge whether all objects in such a domain are syntactic entities like formal proofs.

A QLP *interpretation* on the augmented frame $\langle W, R, \mathcal{D} \rangle$ is a mapping $\mathcal{J}$ which interprets the proof operations $\cdot, !, +$ and $\langle \cdot \forall \cdot \rangle$ as follows:

- $\mathcal{J}$ assigns to each primitive n -ary proof term $a(x_1, \dots, x_n)$ a mapping $a^{\mathcal{J}} : \mathcal{D}^n \rightarrow \mathcal{D}$.
- $\mathcal{J}$ assigns to $!$ a mapping $!^{\mathcal{J}} : \mathcal{D} \rightarrow \mathcal{D}$
- $\mathcal{J}$ assigns to $\cdot$ a binary operation $\cdot^{\mathcal{J}} : \mathcal{D} \times \mathcal{D} \rightarrow \mathcal{D}$
- $\mathcal{J}$ assigns to $+$ a binary operation $+^{\mathcal{J}} : \mathcal{D} \times \mathcal{D} \rightarrow \mathcal{D}$
- $\mathcal{J}$ assigns to $\langle \cdot \forall \cdot \rangle$ a mapping $\forall^{\mathcal{J}}$ from the function space of $\mathcal{D}$ to $\mathcal{D}$ itself, i.e. $\forall^{\mathcal{J}} : (\mathcal{D} \rightarrow \mathcal{D}) \rightarrow \mathcal{D}$

An interpretation is also required to meet several closure conditions: 1) for each $w \in W$, $\mathcal{D}(w)$ is closed with respect to $a^{\mathcal{J}}$ for all primitive proof terms a ; 2) similarly, w is closed with respect to $\cdot^{\mathcal{J}}, !^{\mathcal{J}}, +^{\mathcal{J}}$. Another closure condition for $\forall^{\mathcal{J}}$ will be stated below.

Suppose we have an augmented frame $\langle W, R, \mathcal{D}, \mathcal{J} \rangle$. A *valuation* is a mapping from proof variables to members of $\overline{\mathcal{D}}$. It is not required that $v(x)$ be in $\mathcal{D}(w)$ for all $w \in W$. As usual, w is an x -variant of v if $w(y) = v(y)$ for all variables except possibly x . I will use the notation $v[x/r]$ to denote

the x -variant of v that maps x to r . We may now define the denotation of an arbitrary proof term as follows:

- $a^v = a^J$ for a a primitive proof term
- $x^v = v(x)$ for x a variable
- $(t \cdot u)^v = (t^v \cdot^J u^v)$
- $(t + u)^v = (t^v +^J u^v)$
- $(!t)^v = !^J(t^v)$
- Suppose t^w has been defined for all w . For each x , define a mapping, $\langle \lambda x.t \rangle^v : \overline{\mathcal{D}} \rightarrow \overline{\mathcal{D}}$ as follows: for each $r \in \overline{\mathcal{D}}$, $\langle \lambda x.t \rangle^v(r) := t^{v[x/r]}$. We now define $\langle t \forall x \rangle^v = \forall^J(\langle \lambda x.t \rangle^v)$.

The final condition on interpretations can now be stated as follows: if t is a proof term such that for all variables y other than x , $v(y) \in \mathcal{D}(w)$, $\langle t \forall x \rangle^v \in \mathcal{D}(w)$.

Note that we have allowed for cases in which a given valuation v may assign reasons not in $\mathcal{D}(w)$ to some of the variables occurring in a proof term t . It then follows that it is not possible to interpret t as denoting a reason at w . When this is not the case – i.e. when $v(x) \in \mathcal{D}(w)$ for all $x \in \text{FV}(t)$

– we say that t is *meaningful* at w . The same terminology carries over to formulas in the obvious way.

An *evidence function* $\mathcal{E}$ is a mapping that assigns to each $w \in \overline{\mathcal{D}}$, $r \in \mathcal{D}$ and valuation v a set $\mathcal{E}(w, r, v)$ of QLP formulas. The range of this function $\mathcal{E}$ is a set of formulas which may be thought of as those which the reason r provides (not necessarily decisive) evidence for at w when v is used to interpret its free variables. An evidence function is stipulated to satisfy the following closure conditions:

- If $r \notin \mathcal{D}(w)$, then $\mathcal{E}(w, r, v) = \emptyset$ for all v .
- If $R(w_1, w_2)$, then $\mathcal{E}(w_1, r, v) \subseteq \mathcal{E}(w_2, r, v)$ for all r, v . (monotonicity condition)
- If $(\varphi \rightarrow \psi) \in \mathcal{E}(w, r, v)$ and $\psi \in \mathcal{E}(w, s, v)$ then, $\varphi \in \mathcal{E}(w, (r \cdot^J s), v)$. (application condition)
- If $\varphi \in \mathcal{E}(w, r, v)$ and t is any proof term such that $t^v = r$, then $t : \varphi \in \mathcal{E}(w, !^J(r), v)$. (proof checker condition)
- $\mathcal{E}(w, r, v) \cup \mathcal{E}(w, s, v) \subseteq \mathcal{E}(w, (r +^J s), v)$ (sum condition)
- Let t be a proof term and suppose that for all w' such that $R(w, w')$ and $r \in \mathcal{D}(w')$, $\varphi \in \mathcal{E}(w', \langle \lambda x.t \rangle^v, v[x/r])$. Then $(\forall x)\varphi \in \mathcal{E}(w, (t\forall x)^v, v)$.

(uniform verifier condition)

- If the valuations v and v' agree on the free variables of φ , then $\varphi \in \mathcal{E}(w, r, v)$ iff $\varphi \in \mathcal{E}(w, r, v')$.

These conditions can, to some extent, be given heuristic motivations as is done in [31]. For instance, the monotonicity condition enforces the requirement that the property of providing (not necessarily decisive) evidence is not defeasible in the sense that if t^v provides (not necessarily decisive) evidence for φ at w , then it continues to do so at all accessible worlds. The application, proof checker and sum conditions enforce the requirement that the classes of formulas which are evidenced by reasons mirror the relation between proof terms and formulas enforced by the LP axioms. Finally, the uniform verifier condition states that in order for $\langle t\forall x \rangle^v$ to count as evidence for $(\forall x)\varphi(x)$ at w , every substitution instance of $t(x)$ available at all worlds w' accessible from w must count as evidence for $\varphi(x)$.

We may now give the complete definition of a QLP model and the definition of truth in a model:

Definition 2.3.3. A QLP model is a structure $\mathcal{M} = \langle W, R, \mathcal{D}, \mathcal{J}, \mathcal{E}, V \rangle$ such that $\mathcal{J}$ is an interpretation for $\langle W, R, \mathcal{D} \rangle$, $\mathcal{E}$ is an evidence function and V is a propositional valuation. If $\mathcal{M}$ is a model, $w \in W$ and v is an assignment,

we define truth in $\mathcal{M}$ at w relative to v as follows:

- $\mathcal{M}, w \models_v P_i$ iff $w \in V(P)$
- $\mathcal{M}, w \not\models_v \perp$
- $\mathcal{M}, w \models_v \varphi \rightarrow \psi$ iff $\mathcal{M} \not\models_v \varphi$ or $\mathcal{M}, w \models_v \psi$
- $\mathcal{M} \models_v (\forall x)\varphi$ iff $\mathcal{M} \models_{v[x/r]} \varphi$ for all $r \in \mathcal{D}(w)$
- $\mathcal{M}, w \models_v (\exists x)\varphi$ iff $\mathcal{M} \models_{v[x/r]} \varphi$ for some $r \in \mathcal{D}(w)$
- $\mathcal{M}, w \models_v t : \varphi$ iff $\varphi \in \mathcal{E}(w, t^v, v)$ and for all w' such that $R(w, w')$,
 $\mathcal{M}, w' \models_v \varphi$

A formula φ is said to be *true* at w in $\mathcal{M}$ if for all valuations v , $\mathcal{M}, w \models_v \varphi$ and *valid* in $\mathcal{M}$ if $\mathcal{M}, w \models_v \varphi$ for all w and v such that φ is meaningful at w with respect to v . If φ is valid in $\mathcal{M}$, we write $\mathcal{M} \Vdash \varphi$ and carry this notation over to sets of sentences Γ in the standard way – i.e. $\mathcal{M} \Vdash \Gamma$ just in case $\mathcal{M} \Vdash \psi$ for all $\psi \in \Gamma$.

The following less standard terminology will also be applied to QLP models $\mathcal{M} = \langle W, R, \mathcal{D}, \mathcal{I}, \mathcal{E}, V \rangle$:

- 1) $\mathcal{M}$ is said to *meet the primitive term specification* $\mathcal{P}$ just in case for each primitive proof term p , valuation v , and world $w \in W$, we have $\mathcal{P}(p) \subseteq$

$\mathcal{E}(w, p^v, v)$. (This enforces the requirement that if $\mathcal{P}$ assigns a as evidence for the axiom φ , then the evidence function $\mathcal{E}$ treats a^v as a reason for φ at all worlds in $\mathcal{M}$.)

- 2) $\mathcal{M}$ is said to be *fully explanatory* just in case for all formulas φ , worlds w and valuations v , if $\mathcal{M}, w' \Vdash_v \varphi$ for all worlds w' such that $R(w, w')$, then there exists $r \in \mathcal{D}(w)$ such that $\varphi \in \mathcal{E}(w, r, v)$. (Informally speaking, a fully explanatory model is one in which if φ is necessary at w – or on the epistemic interpretation, *knowable* at w – then there is a reason for φ at w .)
- 3) Let $\mathcal{P}$ be a primitive term specification. We say that φ is a *logical consequence of Γ relative to $\mathcal{P}$* and write $\Gamma \Vdash^{\mathcal{P}} \varphi$ if for all models $\mathcal{M}$ meeting $\mathcal{P}$, if $\mathcal{M} \Vdash \Gamma$, then $\mathcal{M} \Vdash \varphi$. Similarly, we say that φ is an *explanatory logical consequence of Γ relative to $\mathcal{P}$* and write $\Gamma \Vdash_e^{\mathcal{P}} \varphi$ if for all fully explanatory models $\mathcal{M}$ meeting $\mathcal{P}$, if $\mathcal{M} \Vdash \Gamma$, then $\mathcal{M} \Vdash \varphi$.

Over the course of the next two subsections, it will be shown that the system QLP is sound with respect to arbitrary models and complete with respect to fully explanatory models. More precisely, the following will be shown:

Proposition 2.3.3. If $\mathcal{P}$ is an axiomatically primitive specification and

$\Gamma \vdash_{\text{QLP}(\mathcal{P})} \varphi$, then $\Gamma \Vdash^{\mathcal{P}} \varphi$.

Proposition 2.3.4. If $\mathcal{P}$ is an axiomatically appropriate primitive and $\Gamma \Vdash_e^{\mathcal{P}} \varphi$, then $\Gamma \vdash_{\text{QLP}(\mathcal{P})} \varphi$.

Soundness

As noted above, the system called **QLP** here is axiomatized differently than the version given by Fitting in [32] (which corresponds to **QLP**₂ as introduced above). However, as we are about to see, the soundness and completeness proofs given there carry over almost completely to the new system.

Recall that **QLP** differs from **QLP**₂ in three respects:

- 1) In **QLP**, **UG** is a basic rather than a derived rule of inference.
- 2) **QLP** replaces **JUG** with **GJUG**.
- 3) **UF** is not an axiom of **QLP**.

The omission of **UF** obviously does not interfere with carrying over the soundness proof from [32]. Hence in order to finish the proof of Proposition 2.3.3, we need only to show the validity of **UG** and **GJUG**.

Proposition 2.3.5. **UG** is a valid rule of inference.

Proof. Suppose $\Gamma \vdash_{\mathbf{QLP}(\mathcal{P})} (\forall x)\varphi(x)$ has been derived in $\mathbf{QLP}$ by UG from $\Gamma \vdash_{\mathbf{QLP}(\mathcal{P})} \varphi(x)$ where $x \notin \text{FV}(\Gamma)$. By IH, $\Gamma \Vdash^{\mathcal{P}} \varphi(x)$. Thus for all $\mathcal{M}, v, w$ such that $\mathcal{M}, w \Vdash_v \Gamma$, if φ is meaningful at w with respect to v , then $\mathcal{M}, w \Vdash_v \varphi(x)$. Now consider an arbitrary valuation v such that $\mathcal{M}, w \Vdash_v \Gamma$ and suppose that $\varphi(x)$ is meaningful at w with respect to v . Since $x \notin \text{FV}(\Gamma)$, we have $\mathcal{M}, w \Vdash_{v[x/r]} \Gamma$ for all $r \in \mathcal{D}(w)$. Hence $\mathcal{M}, w \Vdash_{v[x/r]} \varphi(x)$ for all $r \in \mathcal{D}(w)$. Thus $\mathcal{M}, w \Vdash_v (\forall x)\varphi(x)$ as desired. $\square$

Proposition 2.3.6. GJUG is a valid rule of inference.

Proof. Suppose that $x \notin \text{FV}(\vec{s} : \vec{\Gamma})$ and that $\vec{s} : \vec{\Gamma} \vdash_{\mathbf{QLP}(\mathcal{P})} t(x) : \varphi(x)$. By the induction hypothesis we have that

$$\text{i) } \vec{s} : \vec{\Gamma} \Vdash^{\mathcal{P}} t(x) : \varphi(x).$$

Now suppose that $\mathcal{M}$ meets $\mathcal{P}$ and w_1 and v are such that $\mathcal{M}, w_1 \Vdash_v \vec{s} : \vec{\Gamma}$.

We must show that $\mathcal{M}, w_1 \Vdash_v \langle t(x)\forall x \rangle : (\forall x)\varphi(x)$.

We first claim that for all w_2 such that $R(w_1, w_2)$ and all $r \in \mathcal{D}(w_2)$, $\mathcal{M}, w_2 \Vdash_{v[x/r]} \vec{s} : \vec{\Gamma}$. For since $\mathcal{M}, w_1 \Vdash_v \vec{s} : \vec{\Gamma}$, $\psi_i \in \mathcal{E}(w_1, s^v, v)$ for all $\psi_i \in \Gamma$, it follows that $\psi_i \in \mathcal{E}(w_2, s^v, v)$ by the monotonicity condition on $\mathcal{E}$. Now consider an arbitrary w_3 such that $R(w_2, w_3)$. By the transitivity of R , we have that $R(w_1, w_3)$. Hence since $\mathcal{M}, w_1 \Vdash_v \vec{s} : \vec{\Gamma}$, $\mathcal{M}, w_3 \Vdash_v \psi_i$ for all

$\psi_i \in \Gamma$. Hence $\mathcal{M}, w_2 \Vdash_v \vec{s} : \vec{\Gamma}$. But since $x \notin \text{FV}(\vec{s} : \vec{\Gamma})$, it then follows that $\mathcal{M}, w_2 \Vdash_{v[x/r]} \vec{s} : \vec{\Gamma}$ for all $r \in \mathcal{D}(w_2)$.

Now let w_2 be arbitrary such that $R(w_1, w_2)$. It follows from the foregoing that $\mathcal{M}, w_2 \Vdash_{v[x/r]} \vec{s} : \vec{\Gamma}$ and hence again by i) that $\mathcal{M}, w_2 \Vdash_{v[x/r]} t(x) : \varphi(x)$ for all $r \in \mathcal{D}(w_2)$. It hence follows that

$$\text{ii) } \varphi(x) \in \mathcal{E}(w_2, \langle \lambda x.t \rangle^v(r), v[x/r]) \text{ for all } r \in \mathcal{D}(w_2)$$

By the validity of the reflection axiom, we also have that $\mathcal{M}, w_2 \Vdash_{v[x/r]} \varphi(x)$ for all $r \in \mathcal{D}(w_2)$ and thus also

$$\text{iii) } \mathcal{M}, w_2 \Vdash_v (\forall x) : \varphi(x)$$

Since w_2 was arbitrary, by ii) we have by the uniform verifier condition on $\mathcal{E}$ that

$$\text{iv) } (\forall x)\varphi(x) \in \mathcal{E}(w_1, \langle t\forall x \rangle^v, v)$$

Finally, we have $\mathcal{M}, w_1 \Vdash_v \langle t\forall x \rangle : (\forall x)\varphi(x)$ by iii) and iv). $\square$

Completeness

The completeness proof given in [32] for QLP_2 combines a maximally consistent set construction together with a Henkin construction by which new constants are added to the language of QLP to provide witnesses for existential statements. Using these methods, a structure $\mathcal{M}^* = \langle W^*, R^*, \mathcal{D}^*, \mathcal{J}^*, \mathcal{E}^*, V^* \rangle$

is defined in a manner which is similar to the canonical model construction for a traditional first-order modal logic – e.g. W^* is taken to be the set of QLP maximally consistent sets in the extend language, $\mathcal{D}^*(\Gamma)$ is defined to be the set of closed terms appearing in Γ , etc. It is then claimed that $\mathcal{M}^*$ is indeed a QLP model. The crux of the argument turns on showing that the function $\mathcal{E}^*$ defined syntactically in $\mathcal{M}^*$ satisfies all of the conditions imposed on evidence functions. And the crux of this argument comes down to showing that $\mathcal{E}^*$ satisfies the closure condition for the uniform verifier symbol.

It is precisely at this point (and at this point only) where Fitting’s completeness proof requires that UF is an axiom of QLP_2 . Since we have seen that all instances of UF are derivable in QLP, the completeness of QLP can be established simply by citing this fact at the appropriate point in Fitting’s proof. However, it is also possible to establish that $\mathcal{E}^*$ has the required property by a direct proof which makes use of GJUG and is accordingly somewhat simpler. In order to present this argument, I will now embark on a highly compressed summary of the completeness proof given in [32], primarily for the purposes of fixing terminology and notation.

Let $p_1, p_2, \dots$ be an infinite list of new variables not appearing $\mathcal{L}_{\text{QLP}}$. We call these *parameters*. Define $\mathcal{L}_{\text{QLP}}^*$ to be the language extending $\mathcal{L}_{\text{QLP}}$ in which parameters are allowed to appear in formulas but are never quantified

and never appear in uniform verifier terms. A $\mathcal{L}_{\text{QLP}}^*$ formula or term is said to be *psuedo-closed* if the only free variables it contains are parameters.

If $\mathcal{P}$ is an axiomatically appropriate primitive term assignment for $\mathcal{L}_{\text{QLP}}$ meeting the free variable condition, it can be extended to an axiomatically appropriate primitive term specification $\mathcal{P}^*$ for $\mathcal{L}_{\text{QLP}}^*$ in the obvious manner – i.e. if $\varphi(x_1, \dots, x_n) \in \mathcal{P}(a)$, then put $\varphi(p_1, \dots, p_n) \in \mathcal{P}(a)$ for each sequence of parameters $p_1, \dots, p_n$ (the choice of $x_1, \dots, x_n$ doesn't matter because $\mathcal{P}$ meets the free variable condition). The following may be shown:

Proposition 2.3.7. Suppose $\varphi(p)$ is a formula of $\mathcal{L}_{\text{QLP}}^*$ where p is a parameter and that $\Gamma \vdash_{\text{QLP}(\mathcal{P}^*)} \varphi(p)$. If $x \notin \text{FV}(\Gamma)$, then $\Gamma \vdash_{\text{QLP}(\mathcal{P}^*)} \varphi(x)$.

If P is a set of parameters, we write $\mathcal{L}_{\text{QLP}}^*(P)$ to denote the sublanguage of $\mathcal{L}_{\text{QLP}}^*$ containing only parameters from P . A set Γ of closed formulas of $\mathcal{L}_{\text{QLP}}^*$ is said to be *inconsistent* if for some $\Delta \subseteq \Gamma$, $\Delta \vdash \perp$. Otherwise such a set is *consistent*. Γ is *maximally consistent* if it is consistent but no proper superset of it is and it is $\exists$ -complete with respect to $\mathcal{L}_{\text{QLP}}^*(P)$ if whenever $\neg(\forall x)\varphi(x) \in \Gamma$, $\neg\varphi(t) \in \Gamma$ for some pseudo-closed proof term t over $\mathcal{L}_{\text{QLP}}^*(P)$. The following can now be demonstrated by a standard Henkin construction:

Proposition 2.3.8. Suppose that Γ is a consistent set of closed $\mathcal{L}_{\text{QLP}}^*(P)$ formulas where P omits infinitely many parameters. Then there is an ex-

tension $\Gamma^* \supseteq \Gamma$ such that Γ consists of closed $\mathcal{L}_{\text{QLP}}^*(P)$ (where $P^* \supseteq P$ also omits infinitely many parameters) such that Γ^* is maximally consistent with respect to P^* and $\exists$ -complete.

A set which is maximally consistent and $\exists$ -complete with respect to $\mathcal{L}_{\text{QLP}}^*(P)$ will be called *P-world-like* and just *world like* if it is *P-world-like* for some P .

Before defining the canonical model, a syntactic device must be introduced which allows us to connect valuations and the replacement of free variables with psuedo-closed terms. To this end, define a *substitution* σ to be a function mapping proof variables *other than parameters* to psuedo-closed terms. We write $t\sigma$ to denote the result of replacing free occurrences of variables x in the domain of σ with $\sigma(x)$. $\varphi\sigma$ is defined similarly for QLP formulas. As the domain of the canonical model $\mathcal{M}^*$ will consist of the set $\overline{\mathcal{D}^*}$ of pseudo-closed terms over $\mathcal{L}_{\text{QLP}}^*$, there is a natural connection between valuations in $\mathcal{M}^*$ and substitutions. In particular, suppose that v is valuation mapping variables of $\mathcal{L}_{\text{QLP}}$ into $\overline{\mathcal{D}^*}$. In this case we write σ_v for the substitution such that $\sigma_v(x) = v(x)$. It may be shown that for all proof term t over $\mathcal{L}_{\text{QLP}}^*$ that $t^v = t\sigma_v$.

Let $\mathcal{P}$ be the the primitive term specification mentioned in the statement of Proposition 2.3.4. The structure model $\mathcal{M}^* = \langle W^*, R^*, \mathcal{D}^*, \mathcal{J}^*, \mathcal{E}^*, V^* \rangle$ is

now defined as follows:

- W^* is the set of world-like sets of $\mathcal{L}_{\text{QLP}}^*(P)$ where the notion of maximal consistency is defined relative to $\text{QLP}(\mathcal{P}^*)$.
- For $\Gamma \in W^*$, define $\Gamma^\sharp = \{\varphi \mid t : \varphi \in \Gamma \text{ for some } t\}$. Now let $\Delta, \Gamma \in W^*$ be such that Γ is P -world-like and Δ is Q -world-like. We then put $R^*(\Gamma, \Delta)$ iff $\Gamma^\sharp \subseteq \Delta$ and $P \subseteq Q$.
- If Γ is P -world-like, then $\mathcal{D}^*(\Gamma)$ is the set of all pseudo-closed proof terms over $\mathcal{L}_{\text{QLP}}^*(P)$.
- $(t)^\mathcal{J}$ is defined to be the identity map if t is a primitive proof term or is of the form $!t_1$, $t_1 \cdot t_2$, or $t_1 + t_2$. In order to define $\forall^\mathcal{J}$, let t be an arbitrary term of $\mathcal{L}_{\text{QLP}}^*$ whose only free variables are parameters except possibly x . We define the mapping $\langle \lambda^* x. t(r) \rangle = t\sigma[x/r]$ for all $r \in \overline{\mathcal{D}^*}$ where $\sigma[x/r]$ is the substitution which replaces x with r (note that r is always free for x since t is pseudo-closed and parameters are never quantified). We now define $(\langle t\forall x \rangle)^\mathcal{J} = \langle \lambda^* x. t \rangle$.
- For all closed formulas of φ of $\mathcal{L}_{\text{QLP}}$, $r \in \overline{\mathcal{D}^*}$ and valuations v , $\varphi \in \mathcal{E}^*(\Gamma, r, v)$ iff $r : (\varphi\sigma_v) \in \Gamma$.
- $V(P) = \{\Gamma \in W^* \mid P \in \Gamma\}$

Proposition 2.3.9. $\mathcal{M}^*$ is a QLP model meeting $\mathcal{P}$.

Proof. It is routine to check that R^* is reflexive and transitive, that $(\cdot)^J$ has the appropriate closure properties and that $\mathcal{M}^*$ meets $\mathcal{P}$. It is also easy to show that $\mathcal{E}^*$ satisfies all the properties on evidence functions other than the uniform verifier condition. (In fact all of these arguments carry over directly from the canonical models of LP and QLP₂ as given in [31] and [32].)

In order to verify that $\mathcal{E}^*$ satisfies the uniform verifier condition, suppose that $\Gamma \in W^*$ is a P -world-like maximally consistent set and suppose that φ is such that $\mathcal{E}$ satisfies the following property corresponding to the left-hand side of the uniform verifier condition:

i) for all QLP proof terms t and formulas φ over $\mathcal{L}_{\text{QLP}}^*$,

$$\varphi(x) \in \mathcal{E}(\Delta, \langle \lambda x.t \rangle^v(r), v[x/r]) \text{ for all } \Delta \text{ such that } R(\Gamma, \Delta) \text{ and } r \in \mathcal{D}(\Delta)$$

We must show that $\mathcal{E}$ satisfies the right-hand side of the uniform verifier condition – i.e.

ii) $(\forall x)\varphi(x) \in \mathcal{E}(\Gamma, \langle t\forall x \rangle^v, v)$

We claim that if i) holds, then for some set of $\mathcal{L}_{\text{QLP}}^*$ formulas of the form $\vec{s} : \vec{\Gamma} = \{s_1 : \psi_1, s_2 : \psi_2, \dots\}$ such that $x \notin \text{FV}(\vec{s} : \vec{\Gamma})$ and $\vec{s} : \vec{\Gamma} \subseteq \Gamma$ the following holds:

iii) $\vec{s} : \vec{\Gamma} \vdash_{\text{QLP}(\mathcal{P}^*)} t(x) : \varphi(x)$

For suppose i) holds and $\vec{s} : \vec{\Gamma} \not\vdash_{\text{QLP}(\mathcal{P}^*)} t(x) : \varphi(x)$ for all such sets. It then follows that $\Gamma^\sharp \cup \{\neg t(p) : \varphi(p)\}$ is consistent for p a parameter not appearing in $\vec{s} : \vec{\Gamma}$. For since $\vec{s} : \vec{\Gamma} \vdash_{\text{QLP}(\mathcal{P}^*)} \psi_i$ for all $\psi_i \in \Gamma^\sharp$, it follows that if $\Gamma^\sharp \vdash_{\text{QLP}(\mathcal{P}^*)} t(p) : \varphi(p)$, we would also have $\vec{s} : \vec{\Gamma} \vdash t(p) : \varphi(p)$. But then by Proposition 2.3.7, we would also have $\vec{s} : \vec{\Gamma} \vdash_{\text{QLP}(\mathcal{P}^*)} t(x) : \varphi(x)$, in contradiction to our assumption.

Now by Proposition 2.3.8, let Δ be a world-like set extending $\Gamma^\sharp \cup \{\neg t(p) : \varphi(p)\}$. Note that $R^*(\Gamma, \Delta)$ by construction. But since $\neg t(p) : \varphi(p) \in \Delta$, it follows that $\varphi(x) \notin \mathcal{E}(\Delta, \langle \lambda x. t(x) \rangle^v(p), v[x/p])$. But since this contradicts i), iii) must hold.

But now note that since $\vec{s} : \vec{\Gamma} \subseteq \Gamma$, and Γ is world-like, the formulas in $\vec{s} : \vec{\Gamma}$ are pseudo-closed. In particular, if $s : \psi \in \vec{s} : \vec{\Gamma}$, then $x \notin \text{FV}(s : \psi)$. It hence follows from ii) that $\vec{s} : \vec{\Gamma} \vdash_{\text{QLP}(\mathcal{P}^*)} \langle t(x) \forall x \rangle : (\forall x) : \varphi(x)$ by GJUG. Hence by $\langle t(x) \forall x \rangle : (\forall x) \varphi(x) \in \Gamma$ and thus $(\forall x) \varphi(x) \in \mathcal{E}(\Gamma, \langle t(x) \forall x \rangle^v, v)$ as desired. $\square$

Having now established that $\mathcal{M}^*$ is a QLP model meeting $\mathcal{P}$, it is now straightforward to prove the following:

Proposition 2.3.10 (Truth Lemma). Let $\mathcal{M}^*$ be the canonical model as

defined above. For each $\Gamma \in W^*$ and formula φ over $\mathcal{L}_{\text{QLP}}$, and valuation v , if φ is meaningful at w with respect to v , then $\mathcal{M}, \Gamma \Vdash_v \varphi$ if and only if $\varphi\sigma_v \in \Gamma$.

We can also establish the following fact about $\mathcal{M}^*$, the proof of which again carries over from [32]:

Proposition 2.3.11. $\mathcal{M}^*$ satisfies the fully explanatory condition.

The completeness of QLP with respect to fully explanatory models can be demonstrated by arguing by contraposition in the standard way: i) if $\Gamma \not\vdash_{\text{QLP}(\mathcal{P})} \varphi$, then $\Gamma \cup \{\neg\varphi\}$ is consistent (we may assume φ is closed since we have both UG and QLP1); ii) by Proposition 2.3.8, $\Gamma \cup \{\neg\varphi\}$ can be extended to a world-like set Δ ; iii) by Proposition 2.3.10, $\mathcal{M}^*, \Gamma \Vdash_v^{\mathcal{P}} \Gamma$ and $\mathcal{M}, \Gamma \not\vdash_v^{\mathcal{P}} \varphi$ for all valuations v ; iv) since $\mathcal{M}^*$ is also fully explanatory by Proposition 2.3.11, it follows that $\Gamma \not\vdash_e^{\mathcal{P}} \varphi$.

Note finally that the Completeness Theorem for QLP allows us to establish the following:

Proposition 2.3.12. For all sets of $\mathcal{L}_{\text{QLP}}$ Γ and single formulas φ , $\Gamma \vdash_{\text{QLP}} \varphi$ if and only if $\Gamma \vdash_{\text{QLP}_2} \varphi$.

Proof. The right-to-left direction of this claim was already demonstrated above. To show the left to right direction, suppose that $\Gamma \not\vdash_{\text{QLP}_2} \varphi$ and let

$\Delta \supseteq \Gamma \cup \{\neg\varphi\}$ be a world-like set in the sense of QLP_2 . Since the canonical model $\mathcal{M}^*$ just described satisfies both QLP and QLP_2 , we hence have $\Gamma \not\models \varphi$ in the sense of QLP . Hence by the Soundness Theorem for QLP , $\Gamma \not\vdash_{\text{QLP}} \varphi$. $\square$

2.3.4 Relationship with S4

It is also possible to define an embedding between **S4** and QLP which is similar to the mapping $(\cdot)^\circ$ between **S4** and **LP** discussed above. To this end, define the mapping $(\cdot)^\exists$ between the language of **S4** and QLP as follows:

- $(P_i)^\exists = P_i$ for P_i a propositional letter
- $(\cdot)^\exists$ commutes with propositional connectives
- $(\Box F)^\exists = (\exists x)x : F^\exists$

Fitting established the following:

Proposition 2.3.13. $\Gamma \vdash_{\text{S4}} \varphi$ if and only if $\Gamma \vdash_{\text{QLP}(\mathcal{P})} \varphi^\exists$ for any axiomatically primitive term specification $\mathcal{P}$.

The left-to-right direction of this result can be demonstrated straightforwardly by showing that the translation of each axiom of **S4** under $(\cdot)^\exists$ is derivable in QLP . The left-to-right direction requires a semantic argument whereby it is shown that it is possible to transform an **S4** counter-model for φ in an QLP counter-model to $\varphi^\exists$.

2.3.5 Barcan- and Converse Barcan-like formulas

As was noted above, Fitting [29] chose to refer to UBF as the Uniform Barcan Formula on the basis of an analogy between its form and that of what I called EBF (for Explicit Barcan Formula) above – i.e.

$$(EBF) \quad (\forall x)(\exists y)y : \varphi(x) \rightarrow (\exists y)y : (\forall x)\varphi(x) \text{ (where } y \text{ does not occur free in } x)$$

This analogy is strengthened in light of the embeddability of QLP into S4 . For note that if we replace each occurrence of $(\exists y)y :$ in EBF with $\Box$ we obtain the original Barcan Formula – i.e.

$$(BF) \quad (\forall x)\Box\varphi(x) \rightarrow \Box(\forall x)\varphi(x)$$

We may similarly note an analogy between the Converse Barcan Formula

$$(CBF) \quad \Box(\forall x)\varphi(x) \rightarrow (\forall x)\Box\varphi(x)$$

and the QLP formula

$$(ECBF) \quad (\exists y)y : (\forall x)\varphi \rightarrow (\forall x)(\exists y)y : \varphi(x)$$

Recall from [33] that a first-order *varying domain augmented frame* is a structure $\mathcal{F} = \langle W, R, \mathcal{D} \rangle$ which $\langle W, R \rangle$ is a frame and $\mathcal{D}$ is a domain function on W . Such a frame is *monotonic* if for all $w_1, w_2 \in W$ such that $R(w_1, w_2)$

$\mathcal{D}(w_1) \subseteq \mathcal{D}(w_2)$ and *anti-monotonic* if for all such pairs $\mathcal{D}(w_2) \subseteq \mathcal{D}(w_1)$.

The following results concerning BF and CBF were all shown by Kripke [59].

- (2.17) i) CBF is derivable in any axiom system for first-order modal logic containing the K axiom, Universal Generalization and the Necessitation rules which is formulated so that universally quantified variables may be instantiated with free variables.
- ii) A varying domain augmented frame is monotonic if and only CBF is valid in every first-order model based on it.
- iii) A varying domain augmented frame is anti-monotonic if and only BF is valid in every first-order model based on it.

As we will now see, analogs of each of these hold for QLP.

Proposition 2.3.14. ECBF is derivable in QLP.

Proof. Reasoning in QLP, we may parallel Kripke's derivation of CBF in first-order modal logic as follows:

- | | |
|---|---------------------|
| i) $(\forall x)\varphi(x) \rightarrow \varphi(x)$ | QLP1 |
| ii) $i : [(\forall x)\varphi(x) \rightarrow \varphi(x)]$ | Axiom Necessitation |
| iii) $y : (\forall x)\varphi(x) \rightarrow i \cdot y : \varphi(x)$ | LP2, ii) |

- | | |
|--|---------------|
| iv) $i \cdot y : \varphi(x) \rightarrow (\exists y)y : \varphi(x)$ | QLP2 |
| v) $y : (\forall x)\varphi(x) \rightarrow (\exists y)y : \varphi(x)$ | iii), iv) |
| vi) $y : (\forall x)\varphi(x) \rightarrow (\forall x)(\exists y)y : \varphi(x)$ | v), UG, QLP3 |
| vii) $(\exists y)y : (\forall x)\varphi(x) \rightarrow (\forall x)(\exists y)y : \varphi(x)$ | vi), UG, QLP4 |

□

It follows by the Soundness Theorem that ECBF is valid in all QLP models. Of course this should not be surprising since such models are stipulated to satisfy the analog of the first-order monotonicity condition. Call a QLP model *constant domain* if all of its worlds have the same domain. Since constant domain models are *ipso facto* anti-monotonic, it should also not be surprising that a relationship can be demonstrated between such models and the explicit version of BF. Using different evidence closure conditions for the uniform verifier symbol, Fitting [30] showed that UBF is valid in all constant domain QLP₁ models. A variant of this argument yields the following:

Proposition 2.3.15. UBF is valid in all constant domain QLP models.

As mentioned above, UBF can be thought of as a Skolemized version of EBF in which the proof of $\varphi(x)$ promised by the antecedent of EBF is given uniformly by a proof term $t(x)$ for all x . Since QLP does not contain

machinery which allows us to prove the existence of such a $t(x)$ combinatorily, in general there is no reason to think that the theory implies that such a $t(x)$ must exist even if $(\forall x)(\exists y)y : \varphi(x)$ holds. However, we do have the following:

Proposition 2.3.16. EBF is valid in all constant domain fully explanatory QLP models.

Proof. Suppose that $\mathcal{M}$ is a fully explanatory constant domain QLP model, that $y \notin \text{FV}(\varphi(x))$ and that $\mathcal{M}, w_1 \Vdash_v (\forall x)(\exists y)y : \varphi(x)$. Then for all $r \in \mathcal{D}(w_1)$, there is an $s \in \mathcal{D}(w_1)$ such that $\mathcal{M}, w_1 \Vdash_{v[x/r][y/s]} y : \varphi(x)$. From this it follows that for all w_2 such that $R(w_1, w_2)$, $\mathcal{M}, w_2 \Vdash_{v[x/r][y/s]} \varphi(x)$ for all $r \in \mathcal{D}(w_1)$. It follows from the fact that $\mathcal{M}$ is a constant domain and that $y \notin \text{FV}(\varphi(x))$ that $\mathcal{M}, w_2 \Vdash_{v[x/r]} \varphi(x)$ for all $r \in \mathcal{D}(w_2)$. Hence $\mathcal{M}, w_2 \Vdash_v (\forall x)\varphi(x)$. Since w_2 was arbitrary, the fully explanatory condition now gives $(\forall x)\varphi(x) \in \mathcal{E}(w_1, r, v)$ for some $r \in \mathcal{D}(w_1)$. Hence $\mathcal{M}, w_1 \Vdash_{v[x/r]} y : (\forall x)\varphi(x)$ from which it follows that $\mathcal{M}, w_1 \Vdash_v (\exists y)y : (\forall x)\varphi(x)$. $\square$

Fix an axiomatically appropriate primitive term specification $\mathcal{P}$ meeting the free variable condition. An *augmented QLP $\mathcal{P}$ -frame* is a structure $\mathcal{F} = \langle W, R, \mathcal{D}, \mathcal{J} \rangle$ such that W is a class of worlds, R is a transitive and reflexive relation on W and $\mathcal{D}$ is a domain function such that $\mathcal{D}(w)$ contains all the primitive proof terms assigned by $\mathcal{P}$ for all $w \in W$ and $\mathcal{J}$ is a QLP

interpretation relative to $\langle W, R, \mathcal{D} \rangle$. In parallel to the right to left direction of (2.17ii) we have the following:

Proposition 2.3.17. A QLP augmented $\mathcal{P}$ frame is anti-monotonic if all instances of EBF are valid in it.

Proof. Let $\mathcal{F} = \langle W, R, \mathcal{D}, \mathcal{J} \rangle$ be an augmented $\mathcal{P}$ -frame which is *not* anti-monotonic and let $w_1, w_2 \in W$ be such that $R(w_1, w_2)$ and $\mathcal{D}(w_1) \subsetneq \mathcal{D}(w_2)$. We construct a QLP model $\mathcal{M} = \langle W, R, \mathcal{D}, \mathcal{J}, \mathcal{E}, V \rangle$ based on $\mathcal{F}$ that the following instance of EBF is not valid in $\mathcal{M}$:

$$(\forall x)(\exists y)y : (x : P) \rightarrow (\exists y)y : (\forall x)x : P$$

Let $V(P) = W$ and $V(Q)$ be arbitrary for all other propositional letters. Let $\mathcal{E}$ be the minimal evidence function satisfying the following conditions:

- i) $P \in \mathcal{E}(w_1, r, v[x/r])$ for all $r \in \mathcal{D}(w_1)$
- ii) $x : P \in \mathcal{E}(w_1, s, v[x/r][s/y])$ for some fixed $s \in \mathcal{D}(w_1)$ and all $r \in \mathcal{D}(w_1)$
- iii) $x : P \notin \mathcal{E}(w_2, p, v[x/p])$ for some fixed $p \in \mathcal{D}(w_2) \setminus \mathcal{D}(w_1)$

It now follows that $\mathcal{M}, w_1 \Vdash_v (\forall x)(\exists y)y : (x : P)$ for all v . But iii) ensures that $\mathcal{M}, w_2 \not\Vdash_{v[x/p]} x : P$ and hence $\mathcal{M}, w_2 \not\Vdash_v (\forall x)x : P$. It hence follows that for all $q \in \mathcal{D}(w_2)$ $\mathcal{M}, w_1 \not\Vdash_{v[y/q]} y : (\forall x)x : P$. Hence $\mathcal{M}, w_1 \not\Vdash_v (\exists y)y : (\forall x)x : P$. □

Before leaving the topic of Barcan- and Converse Barcan-like formulas, it will be useful to record several observations relating the foregoing results with the intended interpretation of QLP. First recall that following Kripke [59], it is usually acknowledged that there is no interpretation-independent reason to accept either BF or CBF as validities of first-order modal logic. For instance, McGee [72] suggests that the BF should be rejected when $\Box$ is interpreted as metaphysical necessity, but accepted when it is interpreted as a truth operator. And as is well known, CBF is valid when $\Box$ is interpreted as a formal provability operator for PA while BF is not (cf. [7]).

It is thus also natural to ask whether we ought to accept either EBF or ECBF as validities of QLP relative to the various notions of justification discussed in Chapter 1. We may note first that ECBF *does* appear to be a valid principle about justification on each of the interpretations. For if y denotes either an everyday justification or an informal proof of a universal statement of the form $(\forall x)\varphi(x)$, it presumably follows that for every object in the range of this quantifier we can construct a proof that this object satisfies the corresponding instance of $\varphi(x)$ by following the same reasoning used in the QLP derivation given in Proposition 2.3.14. ECBF also appears to remain valid on the formal proof interpretation. For if we suppose that y denotes a proof in PA of $(\forall x)\varphi(x)$ in the language of arithmetic, it is not

hard to see that a version of the same argument can be formalized within PA to derive the fact that every numerical instance $\varphi(\overline{n})$ must also be provable. The derivability of ECBF in QLP thus does not seem to run counter to our intended interpretation of proof quantifiers.

The situation is quite different with respect to EBF. I have already attempted to make an intuitive case against this statement regarded as a principle about informal proofs. We will see in the next chapter that this argument transfers to the arithmetic setting and in this context it also serves to impugn UBF and possibly even GJUG as well. However, the results of this section help to clarify what is required by a relational semantics for QLP if we wish to avoid characterizing either of these principles as validities in the technical sense. In particular, Proposition 2.3.15 shows that if we wish to *invalidate* UBF we need to allow for models with (strictly) monotonic domains, which in turn means allowing for the possibility that the class of evidential objects may be larger in worlds epistemically indistinguishable to our own.¹⁴ This in turn suggests that we should not adopt a general view about proof or justification according to which all evidentiary objects of the appropriate sort are *namable* using expressions from a fixed signature. Given the various problems attendant to providing a general characterization of justification discussed in Chapter 1, it may not seem as if we have particularly

strong grounds to either accept or reject this conclusion. However, I will offer some considerations in favor of this view when we consider the possibility of providing an arithmetic semantics for $\mathcal{L}_{\text{QLP}}$ in the next chapter.

¹⁴The situation with respect to EBF is more complex. For on the one hand, it is easy to see that we can invalidate EBF in a constant domain model as long as it is not required to be fully explanatory. But recall that the canonical model of QLP is fully explanatory. Hence if we restricted our attention to constant domain models, we would then need to adopt EBF as an axiom if we wished to retain a complete axiomatization.

Chapter 3

QLP and the Paradox of the Knower

3.1 Introduction

The result now conventionally referred to as the Paradox of the Knower was first explicitly formulated by Montague and Kaplan [M&K] [76]. As I discussed briefly in Chapter 1, the paradox can be taken to reflect a potential inconsistency between the factivity of a knowledge operator K (conceived as a predicate of sentences) and the existence of self-referential sentences which are equivalent (e.g.) to their own unknowability. Although the existence of sentences has not been taken as a central feature of the notion knowledge within epistemic logic, the treatment of self-reference has, of course, been a central concern to the study of truth since at least the time of Tarski [102]. One moral which has been drawn both from [76] and also from Montague's

follow up paper [75] is that problems pertaining to self-reference may other infect operator-like notions such as logical necessity. As we will see below, the latter notion possesses logical properties which are similar to those of the justification-based notion of knowledge discussed in Chapter 1. It is hence reasonable to say that the Knower Paradox stands at the intersection of two traditions: one which seeks to model reasoning about knowledge by analyzing “knows that” as a propositional operator and another which seeks to model reasoning about truth by treating “is true” as a predicate of sentences.

The goals of this chapter are fourfold: 1) to argue that the notion of knowledge which is relevant to the conceptual setting of the Knower Paradox is that of justification-based knowledge (i.e. knowledge requiring proof) discussed in Chapter 1; 2) to illustrate how the reasoning of the paradox can be naturally reconstructed within QLP if we analyze K as expressing existential quantification over justifications; 3) to show that the paradox depends on an additional logical principle (corresponding to the rule GJUG discussed in Chapter 2) which is not traditionally recognized as a premise of the Knower; and 4) to argue that this rule is suspect relative to the intellectual context in which the paradox is typically framed. On the basis of these considerations, I will tentatively advance the thesis that not only is QLP a useful tool studying the Knower, but that one of the principles on which it is based lies

close to the heart of the paradox. However, the route to this conclusion is complicated by the diversity of distinct developments which were inspired by Montague's [75] generalization of the paradox. In order to keep track of some of the issues which have arisen in this context, it will thus be useful to begin by recounting the way in which the Kowner Paradox was originally presented.

Montague and Kaplan [M&K] originally presented the derivation of the Paradox of Kowner as a special case of the so-called Hangman Paradox.¹ This paradox purports to show that a judge sentencing a prisoner i cannot consistently issue a decree expressing that i will be hanged on one of the mornings of a finite number of days $d_1, \dots, d_n$ but that he will not know until the day of the execution which day it will be. The apparent inconsistency of such a decree derives from the fact that upon hearing such an announcement, i can engage in a backwards induction argument² whereby upon hearing the judge's announcement he can infer that the execution could not take place on any of the days.

¹Terminology in this area is confounded by the fact that the Hangman is now more commonly called the Surprise Exam Paradox. M&K additionally cite several earlier presentations in which the same puzzle was called by different names – e.g. Shaw [95], Weiss [112].

² i can reason as follows: i) the execution cannot occur on d_n , for if it did, I would know on the evening of d_{n-1} that it must occur on d_n ; ii) since the execution cannot occur on d_n , it must occur on one of $d_1, \dots, d_{n-1}$; iii) but it cannot take place on d_{n-1} for if it did, I would know on the evening of d_{n-2} that it must occur on $d_{n-1}, \dots$

M&K show how it is possible to formalize this argument in a first-order language $\mathcal{L}_K$ with a predicate $K(x)$. They construe knowledge as a relation between agents and sentences and accordingly state that this predicate is applicable to names of sentences yielding the intended interpretation

$$(3.1) \quad K(\overline{\varphi}) \quad \rightleftharpoons \quad i \text{ knows that } \varphi$$

M&K consider three ways in which such a naming device can be accommodated in a first-order language: i) by the use of quotation marks; ii) by the use of a syntax language in the sense of Tarski [102]; iii) through the use of gödel numbering $\ulcorner \cdot \urcorner$ by which sentences of $\mathcal{L}_K$ are effectively associated with natural numbers. They opt for third mechanism. Although this option looks innocent among the alternatives, it has the consequence that $\mathcal{L}_K$ must be taken to include an arithmetical sub-language $\mathcal{L}_a = \langle +, \times, \mathbf{s}, 0 \rangle$ and that $K(x)$ should be taken as a predicate applying to natural numbers. This in turn means that the theory T in which we suppose that i reasons must be assumed to include arithmetic axioms sufficient to formalize the reasoning about sentences by reasoning about their representation as gödel numbers.³

These choices set the stage for the eventual introduction of the Knower

³After issuing this proviso, M&K go on to use the notation $\overline{\varphi}$ (which was originally introduced to be neutral between the three options) to denote the numeral representing the gödel number of φ . I'll adopt the more conventional choice of using corner quotes $\ulcorner \varphi \urcorner$ to denote the gödel number of φ and $\overline{\ulcorner \varphi \urcorner}$ to denote the numeral representing this number. When no confusion will result, I will generally omit the overline to improve readability.

Paradox as a “instantaneous” or “zero-day” version of the Hangman Paradox.

M&K first note in the case of the Hangman where $n = 1$ the judge’s decree takes the following form:

(3.2) Unless i knows that this statement is false, i will be hanged tomorrow, but i cannot know on the basis of this decree that he will be hanged tomorrow.

If we assume that what it means for i to know φ “on the basis of” φ is for him to know the material conditional $\varphi \rightarrow \psi$ and introduce the propositional variable to δ_1 to denote this decree, (3.2) can be formalized within $\mathcal{L}_K$ as follows:

$$(3.3) \quad \delta_1 \leftrightarrow (K(\ulcorner \neg \delta_1 \urcorner) \vee (\psi \wedge \neg K(\ulcorner \delta_1 \rightarrow \psi \urcorner)))$$

In the case where $n = 0$, the corresponding decree would take the form

(3.4) i knows that the present decree is false.

If we call this decree δ_0 , (3.4) it can accordingly be formalized in $\mathcal{L}_K$ as

$$(3.5) \quad \delta_0 \leftrightarrow K(\ulcorner \neg \delta_0 \urcorner)$$

³Although M&K do not bother to do so, a paradox can be derived directly from (3.3), together with the principles T, U, I discussed below and the additional premise that each instance of these principles is known. See Egré [25] for a formulation of the relevant derivation.

Before we can attempt to determine the consequences of a statement like (3.2), something must first be said about how the existence of sentences δ_1 and δ_0 satisfying conditions like (3.3) and (3.5) are to be secured. We may first note that the former statements uses the demonstrative “this decree” to achieve reference to the statements themselves. As such, M&K propose to treat these statements as *self-referential*.⁴ In order to do this, they propose that we choose T in such a way that it is sufficiently strong to formalize elementary reasoning about syntax. Although they are not completely explicit about which theory they settle on, it has become conventional to suppose that T extends Robinson arithmetic $\mathbf{Q}$ and is hence strong enough to derive the Gödel Diagonal Lemma – i.e.

Lemma 3.1.1 (Diagonal Lemma ([36])). Let $T \supseteq \mathbf{Q}$ and let $\varphi(x)$ be an open sentence in the language of T . There then exists a sentence δ such that $T \vdash \delta \leftrightarrow \varphi(\ulcorner \delta \urcorner)$.

The existence of δ_1 and δ_0 is thus ensured by citing this result.

With these pieces now in place, the Knower Paradox can be framed as

⁴Independently of M&K Fitch [28] also proposed that statements like (3.2) should be treated as instances of self-reference. Although Kripke [58] argues that self reference is not intrinsic to the Hangman Paradox, the necessity of treating statements like (3.2) self-referential will not be of significance since our concern will be with Knower and not the Hangman. Note in particular that although M&K introduce the Knower essentially as a 0-day version of the Hangman (thereby providing an account of the origin of statements like (3.5)), most subsequent commentators have simply taken for granted that $T \supseteq \mathbf{Q}$, from which the provability of statements like (3.5) follows automatically.

soon as we also state several axioms governing the interpretation of $K(x)$ which are assumed to follow from the interpretation (3.1) (and which I'll hereafter refer to as *epistemic principles* to distinguish them from the logical and arithmetic axioms of T). Among these are the familiar principle that knowledge entails truth, which can be formulated $\mathcal{L}_K$ is as follows:

$$(T) \quad K(\ulcorner \varphi \urcorner) \rightarrow \varphi$$

In addition to this, the original formulation of the Knower requires that we assume that i knows this principle – i.e.

$$(U) \quad K(\ulcorner K(\ulcorner \varphi \urcorner) \rightarrow \varphi \urcorner)$$

Finally, we require that i 's knowledge is closed under derivable consequence in T . M&K propose to formalize this by adding to $\mathcal{L}_K$ another a binary predicate $I(x, y)$ with the intended interpretation “ ψ can be derived from premise φ in T .”⁵ Using this notation, the relevant principle may be formulated in $\mathcal{L}_K$ as follows:

$$(I) \quad (I(\ulcorner \varphi \urcorner, \ulcorner \psi \urcorner) \wedge K(\ulcorner \varphi \urcorner)) \rightarrow K(\ulcorner \psi \urcorner)$$

⁵It is clear that this does not require a genuine extension of the language as we may define this relation over $\mathcal{L}_K$ as $I(x, y) =_{df} \exists z [z = i(\ulcorner \varphi \urcorner, \ulcorner \psi \urcorner) \wedge Prov_T(z)]$ where $i(x, y)$ is the (numeralwise representable) function which maps the godel numbers of φ and ψ to the godel number of $\varphi \rightarrow \psi$ and $Prov_T(x)$ is a standard provability predicate for T as discussed below.

The original formulation of the Kowner Paradox may now be stated as follows:

Proposition 3.1.1. Let $T_1 \supseteq \mathbf{Q}$ be a $\mathcal{L}_K$ be a theory containing all instances of T,U and I. Then T is inconsistent.

Proof. By the Diagonal Lemma, find a sentence δ_0 satisfying (3.5). We may then reason in T_1 as follows:

- i) $T \vdash \delta_0 \leftrightarrow K(\ulcorner \neg \delta_0 \urcorner)$ (3.5)
- ii) $T_1 \vdash K(\ulcorner \neg \delta_0 \urcorner) \rightarrow \neg \delta_0$ T
- iii) $T_1 \vdash \delta_0 \rightarrow \neg \delta_0$ i), ii)
- iv) $T_1 \vdash \neg \delta_0$ iii)
- v) $T_1 \vdash I(\ulcorner K(\ulcorner \neg \delta_0 \urcorner) \rightarrow \neg \delta_0 \urcorner, \ulcorner \neg \delta_0 \urcorner)$ i) - iv)
- vi) $T_1 \vdash K(\ulcorner K(\ulcorner \neg \delta_0 \urcorner) \rightarrow \neg \delta_0 \urcorner)$ U
- vii) $T_1 \vdash [K(\ulcorner K(\ulcorner \neg \delta_0 \urcorner) \rightarrow \neg \delta_0 \urcorner) \wedge I(\ulcorner K(\ulcorner \neg \delta_0 \urcorner) \rightarrow \neg \delta_0 \urcorner, \ulcorner \neg \delta_0 \urcorner)) \rightarrow K(\ulcorner \neg \delta_0 \urcorner)]$ I
- xiii) $T_1 \vdash K(\ulcorner \neg \delta_0 \urcorner)$ v), vi), vii)
- ix) $T_1 \vdash \delta_0$ i), viii)
- x) $T_1 \vdash \perp$ iv), ix)

□

M&K join the consensus view that the Hangman Paradox teaches us that decrees like (3.2) cannot be consistently expressed. However, they offer no such neat resolution to the Knower Paradox. For on the one hand, they put forward the T, U and I as reasonable principles governing the interpretation of $K(x)$ under the interpretation (3.1). In fact, upon noting their mutual inconsistency when taken together with Q, M&K do not propose that any of these principles should be abandoned. Instead, they speculatively propose three potential resolutions: 1) “know that” should not be used in the object language of the theory we adopt for reasoning about an agent’s knowledges; 2) we must adopt a hierarchy of metalanguages in the style of Tarski [102] for formalizing reasoning about knowledge; 3) we should exempt some part of elementary syntax from inclusion within T_1 so that the Diagonal Lemma no longer applies. Of course these alternatives are by no means exhaustive of the options we have in responding to the Knower paradox. In fact, most subsequent commentators on the paradox have responded by challenging one of T,U or I. But before surveying such options for replying to the Knower, it will be useful to consider two additional Knower-like paradoxes which are formulated in Montague’s [75] well known follow-up article to [76].

Proposition 3.1.2. Let $T_2 \supseteq Q$ be a $\mathcal{L}_K$ theory containing all instances of T, U and the following two principles:

$$(K) \quad K(\ulcorner \varphi \rightarrow \psi \urcorner) \rightarrow (K(\ulcorner \varphi \urcorner) \rightarrow K(\ulcorner \psi \urcorner))$$

(Log) $K(\ulcorner \varphi \urcorner)$ if φ is an axiom of first-order logic with identity⁶

Then T_2 is inconsistent.

Proof. Let Q be the conjunction of the axioms of $\mathbf{Q}$ and let $\vdash_L \varphi$ (for φ an $\mathcal{L}_K$ sentence) denote that φ is provable from logical axioms alone (i.e. without arithmetic or epistemic axioms). Note that it follows by a straightforward induction using Log and K that

$$(3.6) \quad \text{If } \vdash_L \varphi, \text{ then } T_2 \vdash K(\ulcorner \varphi \urcorner)$$

Finally, by the Diagonal Lemma, we may find a sentence δ_2 satisfying

$$(3.7) \quad \mathbf{Q} \vdash \delta_2 \leftrightarrow K(\ulcorner Q \rightarrow \neg \delta_2 \urcorner)$$

We may now reason in T_2 as follows:

- i) $\vdash_L Q \rightarrow (\delta_2 \leftrightarrow K(\ulcorner Q \rightarrow \neg \delta_2 \urcorner))$ deduction theorem from (3.7)
- ii) $\vdash_L (K(\ulcorner Q \rightarrow \neg \delta_2 \urcorner) \rightarrow (Q \rightarrow \neg \delta_2)) \rightarrow (Q \rightarrow \neg \delta_2)$ i), prop. logic
- iii) $T_2 \vdash K(\ulcorner K(\ulcorner Q \rightarrow \neg \delta_2 \urcorner) \rightarrow (Q \rightarrow \neg \delta_2) \urcorner) \rightarrow (Q \rightarrow \neg \delta_2)$ ii) by (3.6)
- iv) $T_2 \vdash K(\ulcorner K(\ulcorner Q \rightarrow \neg \delta_2 \urcorner) \rightarrow (Q \rightarrow \neg \delta_2) \urcorner)$ U
- v) $T_2 \vdash K(\ulcorner Q \rightarrow \neg \delta_2 \urcorner)$ iii), iv), K
- vi) $T_2 \vdash \delta_2$ (3.7), v)

- vii) $T_2 \vdash K(\ulcorner Q \rightarrow \neg\delta_2 \urcorner) \rightarrow (Q \rightarrow \neg\delta_2)$ T
- viii) $T_2 \vdash Q \rightarrow \neg\delta_2$ v), vii)
- ix) $T_2 \vdash \neg\delta_2$ from viii) since $\mathbf{Q} \subseteq T$
- x) $T_2 \vdash \perp$ vi) and ix)

□

Proposition 3.1.3. Let $T_3 \supseteq \mathbf{Q}$ be a $\mathcal{L}_K$ theory containing all instances of T and satisfying the rule

(Nec) If $T_3 \vdash \varphi$, then $T_3 \vdash K(\ulcorner \varphi \urcorner)$.

Then T_3 is inconsistent.

⁶We must also require that the axiomatization is given so that modus ponens is the only rule of inference. An axiom system for first-order logic with this property is given by Quine [88].

Proof. By the Diagonal Lemma we may find a sentence δ_3

$$(3.8) \quad T_3 \vdash \delta_3 \leftrightarrow \neg K(\ulcorner \delta_3 \urcorner)$$

We may now reason in T_3 as follows:

- i) $T_3 \vdash \delta_3 \leftrightarrow \neg K(\ulcorner \delta_3 \urcorner)$ (3.8)
- ii) $T_3 \vdash K(\ulcorner \delta_3 \urcorner) \rightarrow \delta_3$ T
- iii) $T_3 \vdash \neg K(\ulcorner \delta_3 \urcorner)$ i), ii)
- iv) $T_3 \vdash \delta_3$ i), iii)
- v) $T_3 \vdash K(\ulcorner \delta_3 \urcorner)$ Nec, v)
- vi) $T_3 \vdash \perp$ iii), iv)

□

As Montague observes, the foregoing derivations are all similar to the Liar Paradox in that their derivation depends on the construction of a self-referential statement. But whereas the derivation of the Liar depends on both directions of the Tarski T-schema – i.e. $Tr(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$, where $Tr(x)$ is a truth predicate for a language containing first-order arithmetic – the various forms of the Kowner show how a contradiction can be derived for a predicate $K(x)$ satisfying only the left-to-right direction of such a schema. However, dropping this half of the schema comes at the expense of having

to assume additional axioms involving $K(x)$ which need not be assumed of a truth predicate in order to derive a contradiction.

In the case of each of the above derivations, it can be seen that the extra axioms are required in order to perform what I will refer to as *internalized* reasoning – i.e. to show that it is possible to carry out an inference derivable in T under the scope of $K(x)$. For instance, from the proof of Proposition 3.1.1, we can see that in order to reach a contradiction, the derivation of $\neg\delta_0$ must be internalized to yield $K(\ulcorner\neg\delta_0\urcorner)$ at step xiii) (which is then combined with fixed-point statement i) to derive δ_0). The fact that T is able to internalize some portion of its theorems is a reasonable property to assume of such a theory on the understanding that it is intended to formalize an agent i 's reasoning about his own epistemic capacities. But since there might be a difference between what we can derive about i from the outside and what he can derive about himself, it is also reasonable to keep track of the axioms and rules i may employ in reasoning about his own capacities.

Propositions 3.1.1, 3.1.2 and 3.1.3 can be seen as reporting that three different sets of plausible principles used for internalization each lead to inconsistency. Propositions 3.1.1 and 3.1.2 assume that the axiom T is internalized, as embodied in the axiom U. In relying on axiom I, Proposition 3.1.1 additionally assumes that i can internalize any statement ψ derivable

in T_1 (including those whose derivation depends on arithmetic or epistemic principles) on the presumption that $T_1 \vdash \varphi \rightarrow \psi$ and that φ is internalizable. Proposition 3.1.2 shows how this requirement can be weakened to the case in which the derivation of $T_2 \vdash \varphi \rightarrow \psi$ is mediated only by logical axioms, as embodied by the principles K and Log. On the other hand, Proposition 3.1.3 demonstrates that it is not necessary to assume either U or a closure principle such as I, K or L if we are prepared to accept the rule Nec which states that i can internalize the reasoning leading to any statement derivable in T_3 (again including those derivations which depend on arithmetic and epistemic principles).

Since U and versions of Log and I can be derived from the Nec rule in the presence of axiom T, Proposition 3.1.3 might be thought of as the weakest of the three results presented.⁷ In what follows, however, I will treat this proposition (or more importantly, its derivation) as the basic form of the Knower Paradox. Part of the motivation for this will become clearer when we examine the prospects for reconstructing the knower in various forms of propositional modal logic. However, at this point we can already observe that both Propositions 3.1.1 and 3.1.2 depend on the axiom U. As several

⁷More precisely, U is trivially derivable in the system of Proposition 3.1.3 by applying the Nec rule to axiom T, whereas Log is just Nec restricted to logical axioms. In the case that $T_3 \vdash \varphi \rightarrow \psi$ and $T_3 \vdash K(\varphi)$, then it follows via T, modus ponens and Nec that $T \vdash K(\psi)$.

commentators have pointed out, accepting U seems to be tantamount to acknowledging that i can reflect upon (and thereby come to know) the factivity of $K(x)$ as expressed by the axiom T. Needless to say, this latter property is central to our everyday conception of knowledge. And as such, accepting U appears to attribute to i a relatively modest capacity for reflecting on his own epistemic capacities. However, the fact that such a principle is not as innocent as it appears can already be observed by noting that from i 's perspective, there will in general be no way of distinguishing a case in which he believes φ on the basis of evidence he takes to be irrefutable (but in reality is not), from one on which he believes it on the basis of truly irrefutable evidence. As such, it seems reasonable to think of U as *less* secure than the other logical and arithmetic principles embodied by any of the theories in question. For this reason it seems doubtful that we can find an argument for accepting U which does not implicitly serve as an argument for accepting the principle Nec used in Proposition 3.1.3 as well.

It may also be noted, however, that since Proposition 3.1.3 depends on fewer explicitly articulated principles, taking it as our paradigm appears to limit our options for responding to the paradox. However, my strategy here will not be to follow other commentators in arguing that any of T, I or U must be rejected. Rather, I will first attempt to reconstruct the Knower in

a manner which highlights a feature of knowledge which I will now argue is implicit in the setting of the paradox and then argue that once this feature is taken into account, it becomes evident that additional principles are involved in the derivation of the paradox.

The relevant feature is simply the familiar observation that in attributing knowledge of a proposition φ to an agent i , we normally mean to convey that i possesses some sort of justification for φ . That such a conception is central to the conception of knowledge in which the Knower is grounded is evident from the fact that the paradox invites us to think of theories like T_1 , T_2 or T_3 not only as media which we can use to reason externally about i 's knowledge, but also as an abstract model of how he himself might reason. This feature is again clearest in the case of T_3 . For note that the acceptability of the rule Nec appears to premised on our willingness to accept that i can come to know φ on the assumption that φ is provable in the theory itself. But note that if this is the case, there also exists a justification in virtue of i can come to know φ – i.e. its formal derivation in T_3 .⁸

As a number of theorists have noted, this justificatory feature of knowl-

⁸Note that although T_1 and T_2 appear to be more conservative than T_3 in attributing knowledge of non-logical (or non-arithmetic) principles to i , the same sort of justification-based account must presumably be given in order to motivate closure principles like I or K. For instance, the rational for concluding $K(\ulcorner\psi\urcorner)$ in the case that $K(\ulcorner\psi\urcorner)$ and $\varphi \rightarrow \psi$ are both provable presumably derives from the observation that if these conditions obtain, then i can deduce ψ himself by composing a proof of φ with a proof of $\varphi \rightarrow \psi$.

edge also highlights the similarity between the epistemic axioms which are employed in the various versions of the Knower and the principles which hold for arithmetic provability predicates. In particular, if we think that the notion of knowledge formalized by these theories is one which depends on proof in a theory T , then it is reasonable to think that we can further analyze $K(\ulcorner \varphi \urcorner)$ in terms of a formal proof predicate $Proof_T(x, y)$ for T as follows

$$(3.9) \quad K(\ulcorner \varphi \urcorner) \iff (\exists x) Proof_T(x, \ulcorner \varphi \urcorner)$$

The benefit of adopting such an analysis is that it makes explicit the quantification over justifications which I have just argued is implicit in the interpretation of the predicate $K(x)$ in the context of the Knower. Note that arithmetic theories $T \supseteq \mathbf{Q}$ have the property of Σ_1 -completeness – i.e. for every true $\mathcal{L}_a$ statement φ of the form $(\exists x)\psi(x)$ (for $\psi(x)$ containing only bounded quantifiers), if φ is true, then $T \vdash \varphi$. Since via godel-numbering the relation “ φ is provable in T ” can be formalized by the Σ_1 predicate $(\exists x)Proof_T(x, \ulcorner \varphi \urcorner)$, it follows that such a theory will satisfy a version of the Nec rule if $K(x)$ is analyzed according to (3.9).

Adopting (3.9) as an analysis of $K(x)$ may initially seem like a reasonable means of making sense of the proposal that knowledge should be analyzed in terms of the existence of justification. However, a consequence of adopting

such an analysis is that we will then be committed to all of the principles involving $K(x)$ which are provable for a particular choice of arithmetical theory T and definition of $Proof_T(x, y)$. Several well known phenomena demonstrate just how sensitive such principles can be to these choices:

- 1) Suppose that $Proof_T(x, y)$ is defined in the manner of the standard Gödel [36] proof predicate (i.e. as formalizing x is the code of a sequence of formulas whose last member is the formula with code y and each of which is either an axiom of T or follows from earlier members by the rules of T) and T is taken to include a relative strong arithmetic theory such as Peano Arithmetic [PA]. In this case, the conventional Hilbert-Bernay's conditions about $Proof_T(x, y)$ will be derivable in T . These will include the arithmetic analogues of the 4 (i.e. transitivity) and L (i.e. Löb) modal axioms. Written according to (3.9) these are as follows:

$$(4) \quad K(\ulcorner \varphi \urcorner) \rightarrow K(\ulcorner K(\ulcorner \varphi \urcorner) \urcorner)$$

$$(L) \quad K(\ulcorner K(\ulcorner \varphi \urcorner) \rightarrow \varphi \urcorner) \rightarrow K(\ulcorner \varphi \urcorner)$$

While 4 is often taken to be a plausible principle about knowledge (i.e. the so-called *positive introspection principle*), such a principle is generally considered to reflect a stronger assumption about knowledge than T

or closure principles like K assumed in the Knower.⁹ With one partial exception ([6]) the Löb axiom is not generally accepted as a plausible principle about knowledge. In particular, as I will discuss further below, it is transparently inconsistent with T and Nec .

- 2) Suppose we choose to base T on the arithmetic theory Q . Such a theory is arguably the weakest setting in which it is possible to carry out the arithmetization of syntax in a manner so that results like the Diagonal Lemma will be provable. However, if we also adopt an analysis like (3.9), then the principles like we must then accept about $K(x)$ will depend quite sensitively to the definition of $Proof_T(x, y)$ and on specific godel numbering on which it is based. On the hand, the consensus view is that Q is not capable of proving either the Second Incompleteness Theorem or Löb's Theorem in their standard forms.¹⁰ As I will discuss further below, it is clear that there is an incompatibility between these principles and the U axiom. Hence there non-derivability in a particular arithmetic theory might initially seem like a *positive* feature. However, it also well known that if $Proof_T(x, y)$ is defined in the standard manner, Q will also not

⁹4 is, however, often assumed to be a valid principle about rational belief. See Thomason [103] and Koons [55] for versions of the Knower Paradox based on a notion of rational belief that satisfies 4 and U but not T . See also note 15 below for a discussion of a sense in which the formulation of Nec presupposes 4.

prove presumably desirable features such as the analogs of closure axioms

K or I if $K(x)$ is analyzed according to (3.9).

- 3) Following Feferman [26], it is possible to construct proof predicates which coincide extensionally with the standard definition of the Gödel proof predicate but which obey different principles at the propositional level. For example suppose that we define $Proof'_T(x, y) =_{df} Proof_T(x, y) \wedge y \neq \ulcorner 0 = 1 \urcorner$ where $Proof_T(x, y)$ is the Gödel proof predicate. If T is consistent, then $(\exists x)Proof'_T(x, y)$ will numeralwise (i.e. extensionally) represent the class of theorems of T . However, if we now define consistency of T in the standard manner (i.e. as the non-derivability of $0 = 1$), then we will trivially be able to prove a consistency statement for T formulated using $Proof'_T(x, y)$.

These problems illustrate the sorts of technical challenges which must be faced if we attempt to provide a specific structural analysis of the intended meaning of $K(x)$ using metamathematical notions formalizable within T .

¹⁰The situation here is quite complex. The traditional reference is to Bezboruah and Sheperdson [5] who demonstrate the a version of the Second Incompleteness Theorem is derivable in Q . However, Q is so weak as to not be able to derive facts about its own provability predicate $Prov_Q(x)$ (when defined in a conventional manner) which seem to be required for us to recognize that it is indeed of a provability by reasoning within the theory itself. As such, Bezboruah and Sheperdson's themselves state that their result "is devoid of any philosophical interest and that in such a weak system [the formula $(\exists x)Prov_Q(\ulcorner 0 = 1 \urcorner)$] cannot be said to express consistency but only an algebraic property which in a stronger system ... could reasonably be said to express the consistency of Q ."

Presumably, however, we do not want to complicate our of the everyday conception of knowledge by introducing idiosyncrasies which arise from the arithmetic (as opposed to epistemic) axioms of such a theory.¹¹ Although such difficulties will resurface below, it possible to partially circumvent them while retaining a justification-based interpretation of knowledge by using the language of explicit modal logic to formulate the epistemic principles on which the Knower is based. Such a formalization can be based on the following analysis of $K(x)$:

$$(3.10) \quad K(\ulcorner \varphi \urcorner) \iff (\exists x)x : \varphi$$

Pursuant to our prior discussion of quantification in QLP in Chapters 1 and 2, the quantifier in (3.10) need not be interpreted as ranging over proofs in any particular formal system. As such, this analysis states that in order to know φ , there must exist some evidentiary object (such an informal proof) which serves to justify φ for i . Adopting such an analysis does not commit us to any particular characterization of the nature of these objects or our ability to formalize their properties in T . As such (3.10) seems to be a considerably more faithful means of understanding the notion of knowledge relevant to the Knower. For on the one hand we have seen that the motivation of

¹¹In fact Montague [75] appears (p. 295) to cite difficulties of this sort as part of his motivation for not providing a structural analysis of $K(x)$, but rather treating it as an unanalyzed predicate satisfying certain propositional-like axioms.

principles like Nec, I or K appear to depend on understanding knowledge as requiring justification in the form of proof. But neither these principles nor the epistemic axioms T and U appear to require that we analyze the proof of arithmetically.

We will see below that the adoption of (3.10) allows for a fairly straightforward reconstruction of the reasoning of Proposition 3.1.3 (and by extension also Propositions 3.1.1 and 3.1.2) in QLP. In this derivation, the Lifting Lemma for QLP (Proposition 2.3.2) plays a role analogous to the Nec rule in the proof of Proposition 3.1.3. However, since necessitation in the strict sense is confined to axioms in QLP, the reconstructed proof must progress through intermediate steps involving quantificational reasoning about justifications which are not explicitly represented in the derivations of any of the version of the Knower we have considered thus far. Below we will see that the crux of such a derivation is the internalization of the QLP analogue of T – i.e.

$$(\mathbf{T}_q) \ (\exists x)x : \varphi \rightarrow \varphi$$

– to yield an appropriate analog of U

$$(\mathbf{U}_q) \ (\exists y)y : (\exists x)x : \varphi \rightarrow \varphi)$$

I will show that in the case where φ corresponds to a propositional letter, the

derivation of $\mathbf{U}_q$ requires the Generalized Justified Universal Generalization rule [GJUG] in order to internalize the proof of $\mathbf{T}_q$ from the standard LP reflection axiom (i.e. $t : \varphi \rightarrow \varphi$). I will then argue that in the context of the Knower, this step is illegitimate for reasons relating both to its conceptual significance and similarity to arithmetic principles such as the omega-rule.

As I indicated above, judging whether a particular consideration constitute a genuine resolution to the Knower Paradox requires a high degree of circumspection. One reason for this is that although M&K describe Proposition 3.1.1) as a paradox about the everyday notion of knowledge, Montague's second paper [75] initiated a tradition of viewing both this and related inconsistency results like Propositions 3.1.2 and 3.1.3 as antinomies pertaining to a range of other operator-like notions – most notably that of *logical necessity* – which can be taken to satisfies the same axioms and rules. For this reason, results like Proposition 3.1.3 have played a role in a variety of traditions in which the predicate $K(x)$ is given different interpretations. My current strategy will be to focus on the literature (which I refer to as the *epistemic tradition*) which treats these results as limitative findings about the everyday concept of knowledge and argue on this basis that there is a conceptually well motivated resolution to the paradox. With this said, however, concerns originating in other traditions will also be relevant to the development below.

It will also be useful to briefly survey the different literatures in which the results presented in this section have featured.

The epistemic tradition

This tradition treats the results surveyed in this section as bearing on everyday epistemic concepts. As I have already mentioned, this is in line with the interpretation M&K originally assigned to 3.1.1. Subsequent work in this tradition falls into two non-exclusive categories: 1) attempts to isolate additional inconsistency results based on axioms applicable either to knowledge or related epistemic notions just as justified belief; 2) attempts to resolve these inconsistency by adducing conceptual considerations which call for the rejection of some or all of the principles on which the corresponding paradoxes are based. Exemplifying the first category is the work of Thomason [103] and Koons [55] both of whom present Knower-like paradoxes based on axioms which they claim characterize a notion of justified belief. In this category also falls the work of Cross [16] who argues that a version of the Knower can be formulated for a weaker notion of knowledge which is not closed under principles like I or K. Cross's results are targeted against Maitzen [67] who argues that the paradox can be resolved by rejecting principle I (meaning that his work falls in the second category). Also in the second category are

the papers by Anderson [1], Tymoczko [106], Égré [25] and Pogliolesi [81]) all of whom argue that principle U ought to be rejected for the everyday concept of knowledge (a conclusion which Cross also joins).

The logical necessity tradition

One of Montague's foils in [75] was Quine [87] who is famous for having argued that modality may only be coherently be regarded as a predicate of sentences. Although M&K originally presented Proposition 3.1.1 as pertaining to the concept of knowledge, the same result, along with 3.1.2 and 3.1.3 are presented in [75] as limitative results about the notion of necessity. This is achieved by simply replacing the predicate $K(x)$ in the statement of these results with a predicate $N(x)$ and arguing the principles on which they are based remain plausible remain plausible if $N(x)$ is interpreted as expressing necessity. Montague hence concludes on the basis of these results that the principles such as the axiom T commonly thought to hold of the operator expressed by 'it is necessary that' cannot be consistently maintained if modality is treated as a predicate of sentences.¹²

Although he does not say so explicitly, Montague's characterization of necessity appears to coincide with a form of what is often called *logical*

¹² "If necessity is treated syntactically . . . as Carnap and Quine have urged, then virtually all of modal logic . . . must be rejected." [75], p. 294.

necessity – i.e. truth in virtue of logical form. One way of analyzing φ is *logically necessary* is as φ is *provable from logically valid axioms*. On the basis of this analysis McGee [72] proposes that $N(x)$ be analyzed in terms of a formal provability predicate $(\exists x)Proof_T(x, \ulcorner \varphi \urcorner)$ for some theory $T \supseteq \mathbf{Q}$. He argues from this that only must the axiom T be rejected as a valid principle about logical necessity, but other schema provable for sufficiently strong T as the Löb axiom $N(\ulcorner N(\ulcorner \varphi \urcorner \rightarrow \varphi) \urcorner) \rightarrow N(\ulcorner \varphi \urcorner)$ should be accepted as valid principles about logical necessity. This tradition has recently been extended by Halbach, Leitgeb and Welch [43], [44] who both argue that there any several advantages of conceiving of necessity as a predicate of sentences, but also present several forms of relational semantics for theories which include both arithmetic and modal predicates.

The Provability logic tradition

The modal properties of the proof predicate of PA and other systems have been intensively under the the rubric of *provability logic*. The fundamental idea of this study is to characterize the class of sentences of propositional modal logic which are valid when the operator $\Box$ is interpreted as expressing formal provability in a arithmetic theory T . This is achieved by defining a mapping $(\cdot)^*$ (known as an *arithmetic interpretation*) from the language $\mathcal{L}_\Box$

to $\mathcal{L}_a$ defined as follows:

- $P^* =$ an arbitrary sentences of $\mathcal{L}_a$
- $(\cdot)^*$ commutes with boolean connectives
- $(\Box\varphi) = (\exists x)Proof_T(x, \ulcorner \varphi^* \urcorner)$ where $Proof_T(x, y)$ is a proof predicate for T .

The early history of this subject overlaps with the work of Montague [75] both in the sense of being inspired by Gödel's [35] use of a modal operator to express a concept of provability and also in the central role played by the Diagonal Lemma (cf. [9]).

Starting in the early 1970s, however provability logic developed in a manner which was influence by different factors than work on logic necessity or the epistemic resolution to the Knower. For one thing, since the desire was not study the properties of intuitive notions like knowledge or necessity, but rather to provide a modal characterization of the Second Incompleteness Theorem and Löb's Theorem. Thus rather than taking a weak arithmetic theory like $\mathbf{Q}$ as the starting point, provability logic has traditionally be most interested in the case where T is sufficient strong to prove the Hilbert-Bernays conditions for $Prove_T(x, y)$ in a straightforward sense. Hence, most work in this area has focused on the case in which $T = \mathbf{PA}$. The guiding question

of much early work in provability logic was to provide a characterization of the class of modal statements which are either provable in **PA** or true in the standard model for all choices of $(\cdot)^*$. These questions was answer in a seminal paper of Solovay [99] where it was shown that these classes corresponded respectively to the modal theories **GL** and **GLS** (cf. [7]), about which more will be said below. Another major topic in provability logic is modal characterization of self reference. In provability logic, this is achieved by the De Jongh-Sambic Fixed Point Theorem to which I will also return below.

The remainder of this chapter is structured as follows. In the next section, I will consider how various forms of the Knower Paradox can be reconstructed both in propositional modal logic and in **QLP**. This will lead to a further examination of the role both of Diagonal Lemma and internalization in the derivations of the results like Proposition 3.1.3. In Section 3, I will then attempt to reconstruct this modal reasoning in **QLP** so as to highlight the role of the GJUG. In Section 4, I will present an arithmetic semantics for a fragment of **QLP** which I will argue sheds light on the significance of GJUG and why it may not be a plausible principle about the everyday concept of justification. Finally, in Section 5 I will refine these points by comparing what can be learned about the Knower from its **QLP** reconstruction to the

sorts of resolutions which have been offered by other theorists in the epistemic tradition.

3.2 The reconstruction of the Knower in modal logic

Before we attempt to reconstruct the Knower in QLP, it will be useful to first compare the arithmetic forms of the paradox which we considered in the previous section with arguments which can be given in propositional modal logic. For if we put aside for the moment Montague's motivation for treating K as a predicate of sentences, it should be clear that much of the reasoning involved in the derivations of Propositions 3.1.1, 3.1.2 and 3.1.3 remains valid if we interpret K as a modal operator applicable to propositions. In order to make such a comparison precise, we might start out with a propositional modal language $\mathcal{L}_\Box$ of the sort discussed in Chapter 1 in which the modality $\Box$ is subject to the intended interpretation

$$(3.11) \quad \Box\varphi \Rightarrow i \text{ knows } \varphi$$

We might then then seek to interpret $\mathcal{L}_\Box$ formulas of the form $\Box\varphi$ as expressing statements which can be analyzed in $\mathcal{L}_K$ via statements of the form $K(\ulcorner\varphi\urcorner)$.

This idea can be implemented using the same notion of an arithmetic interpretation of a modal language as applied to provability logic. In particular, let $(\cdot)^+$ be a mapping from $\mathcal{L}_\square$ to $\mathcal{L}_K$ defined as follows: P^+ is a fixed sentence of $\mathcal{L}_K$ for all propositional letters, $(\cdot)^+$ commutes with propositional connectives and $(\Box\varphi)^+ = K(\ulcorner\varphi^+\urcorner)$ where $\ulcorner\cdot\urcorner$ is some fixed godel numbering of formulas of $\mathcal{L}_K$. Relative to such an interpretation we can see that the epistemic axioms T, U, K used in the various derivations of the Knower considered in Section 1 are the images of the following modal principles:

$$(\mathbf{T}) \quad \Box\varphi \rightarrow \varphi$$

$$(\mathbf{K}) \quad \Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$$

$$(\mathbf{U}) \quad \Box(\Box\varphi \rightarrow \varphi)$$

Relative to (3.11) **T** and **K** correspond to the familiar modal axioms expressing that i 's knowledge is factive and is closed under modus ponens. Although **U** is not standardly singled out as a modal axiom of particular interest, it is also easy to motivate on the epistemic interpretation – e.g. i can presumably come to know that any instance of the reflection principle $\Box\varphi \rightarrow \varphi$ is true simply by considering the meaning of “knows that.”

I have already flagged the fact that it is the arithmetic analog U of **U** which many commentators in the epistemic tradition have flagged as the

crux of the Knewer Paradox. It is notable in this regard that in the context of propositional modal logic, **U** is not typically taken as an axiom precisely because it is derivable in any normal modal system containing the axiom **T**. In particular, any instance of **U** can be obtained by applying the traditional modal necessitation rule

(**Nec**) If $\vdash \varphi$, then $\vdash \Box\varphi$.

to an appropriate instance of **T**. As we have seen, **Nec** can be motivated in the epistemic context based on the idea that proof is sufficient for knowledge – i.e. if φ corresponds to a principle in $\mathcal{L}_\Box$ which is provable by reasoning from accepted logical and epistemic principles, then i can presumably come to φ by constructing its proof. In the propositional setting, the argument for the acceptability of **U** thus runs as follows: i) since we accept **T** as a valid principle about knowledge, it is *ipso facto* provable; ii) on this basis, i can come to know **T** (a fact which in $\mathcal{L}_\Box$ is expressed as $\Box(\Box\varphi \rightarrow \varphi)$) by examining its proof.

Since when presented as a formal derivation this proof is essentially trivial, there seems to be nothing implicit in our desire to assign (3.11) as the interpretation of $\Box$ to prefer a system in which **U** is assigned the status of an axiom from which it is derived from what might be taken as more fun-

damental principles about knowledge such as **T** and **Nec**. We will see in the next section, however, that the proof of the QLP principle $\mathbf{U}_q$ from $\mathbf{T}_q$ is more involved. If there are already reasons to be suspicious of **U**, then the principles on which the proof of $\mathbf{U}_q$ in QLP depends thus also demand scrutiny within the epistemic tradition.

In order to see how **U** figures in the derivation of the Knower, we must first attempt to reconstruct the reasoning of the derivations given in Section 1 in the language $\mathcal{L}_\square$. We have just seen how that **T**, **K** and **U** correspond to principles which can be stated in the language of propositional modal logic. However, the situation is less straightforward for **I**, **Log** and **Nec**. As observed above, each of these principles contributes to the ability of an arithmetic T to internalize reasoning which can be conducted in T . As such, these principle are intended to some part of the role of **Nec** in the modal setting. But it is notable that in a proof theoretic sense, none of the principles corresponds to **Nec** in a completely straightforward manner. For instance, **Log** allows us to conclude $K(\ulcorner \varphi \urcorner)$ only in the case where φ is an axiom of first-order logic, a restriction which cannot even be formulated using notions applicable to $\mathcal{L}_\square$.¹³ And for reasons which we examine in greater detail in a moment,

¹³A more direct comparison can be drawn if, instead of **Nec**, we adopt the rule **Nec_L** which allows only axioms of the modal **L** in which we are working to be necessitated. If **L** also contains **K**, then **Nec** will obviously be a derivable rule of the resulting system. Note, however, **Log** still does not correspond to **Nec_L** as the latter but not the former may be

the principle **Nec** does not directly correspond to the modal rule **Nec** since the former but not the latter allows for the derivation of φ to depend on hypotheses in the forms of the non-logical axioms of T .

Temporarily putting aside the problem of finding an appropriate modal analog of these principles, the other evident problem in attempting to reconstruct any derivations of the Knower given in Section 1 in $\mathcal{L}_\square$ is that there is no reason to expect a modal logic T (which contains the axioms **T**, **K** and the rule **Nec**) proves a statement of the form

$$(3.12) \quad \varphi \leftrightarrow \neg \square \varphi$$

which can be taken as analog of the self-referential statement (3.8) used in the proof of Proposition 3.1.3. This problem is closely related to Montague's attack on Quine's proposal that modality be treated as a predicate of sentences. For as we have seen, Montague argues that the adoption of such a treatment requires that the theory we employ for reasoning about sentential operators such as "knows that" and "is necessary that" include a sufficient amount of arithmetic to formalize reasoning about syntax. Since the Diagonal Lemma is provable in even weak systems like **Q**, the existence of self-referential sentences like (3.8) seems to inevitably follow. But while it is now well known that there is a way in which self-reference can arise within

applied to non-logical epistemic axioms such as **T**.

propositional modal logic, not only was this not yet understood in the 1960s, there is also a straightforward sense in which the existence of states satisfying (3.12) can be seen to be incompatible with modal logics containing T .

In order to see this, we may simply assume an instance of (3.12) and then attempted to reason in T . It is easy to see that for no $\mathcal{L}_\square$ sentence φ will the corresponding instance of (3.12) be provable in T (for if it were, then the same reasoning of Proposition 3.1.3 would yield a contradiction). But in fact, something stronger is true:

Proposition 3.2.1. For all φ , $\mathsf{T} \vdash \neg \square(\varphi \leftrightarrow \neg \square \varphi)$

Proof. Reasoning in T :

- | | |
|---|-------------------------------|
| i) $\vdash \square(\varphi \leftrightarrow \neg \square \varphi) \rightarrow (\square \varphi \rightarrow \neg \varphi)$ | T |
| ii) $\vdash \square(\varphi \leftrightarrow \neg \square \varphi) \rightarrow (\square \varphi \rightarrow \varphi)$ | propositional logic, T |
| iii) $\vdash \square(\varphi \leftrightarrow \neg \square \varphi) \rightarrow \neg \square \varphi$ | propositional logic, i), ii) |
| iv) $\vdash \square(\varphi \leftrightarrow \neg \square \varphi) \rightarrow \square(\square \varphi \rightarrow \neg \varphi)$ | Nec, K |
| v) $\vdash \square \varphi \rightarrow \varphi$ | T |
| vi) $\vdash \square(\square \varphi \rightarrow \varphi)$ | Nec v) |
| vii) $\vdash \square(\varphi \leftrightarrow \neg \square \varphi) \rightarrow \square(\square \varphi \rightarrow \varphi)$ | propositional logic, vi) |
| viii) $\vdash \square(\varphi \leftrightarrow \neg \square \varphi) \rightarrow \square \neg \square \varphi$ | K, iv), vii) |
| ix) $\vdash \square(\varphi \leftrightarrow \neg \square \varphi) \rightarrow (\square \neg \square \varphi \rightarrow \square \varphi)$ | propositional logic, K |

$$\text{x)} \quad \vdash \Box(\varphi \leftrightarrow \neg\Box\varphi) \rightarrow \Box\varphi \quad \text{viii), xi)}$$

$$\text{xi)} \quad \vdash \neg\Box(\varphi \leftrightarrow \neg\Box\varphi) \quad \text{iii), x)}$$

□

Relative to the interpretation (3.11), Proposition 3.2.1 can be taken to demonstrate that not only can *i* not come to know a statement of the form $\varphi \leftrightarrow \neg\Box\varphi$ to be true, but in fact he can prove that he cannot know such a statement. This observation should be contrasted with the significance typically assigned to results like Proposition 3.1.3 – i.e. that otherwise plausible seeming epistemic principles like **T** and **Nec** are inconsistent with the arithmetic principles required for the formalization of syntax. As we have seen in Chapter 1, however, independent motivation can be adduced for regarding a propositional modal systems such as **S4** which extend **T** as media for reasoning about informal provability. As I will see below, this notion appears to be closely related to the notion of proof-based knowledge on which I argued in the previous section the Knower is based. We might thus equally well view Proposition 3.2.1 as demonstrating the inadmissibility of the informal provability of the existence of certain kinds of self-referential statements containing the informal provability operator.

Proposition 3.2.1 parallels Proposition 3.1.3 in the sense that if we assume the provability of a statement of the form (3.12), then a contradiction may be derived from the modal principles **T** and **Nec**. However, it is also notable that the modal derivation just given does not mirror that of Proposition 3.1.3. One difference is that in the case of Proposition 3.2.1, the statement $\Box(\varphi \leftrightarrow \neg\Box\varphi)$ must be carried along as the antecedent to a conditional since it is not provable in **T** itself. This is required because if we assumed $\Box(\varphi \leftrightarrow \neg\Box\varphi)$ as a hypothesis to the left of the turnstile, then the application of **Nec** at steps iv) and vi) would not be licensed. It is notable, however, that if we adopt a form of necessitation rule which allows for hypotheses such as the **SNec** introduced in Chapter 2 (i.e. $\Box\Gamma \vdash \varphi \therefore \Box\Gamma \vdash \Box\varphi$ where all $\psi \in \Box\Gamma$ are of the form $\psi \equiv \Box\chi$), then the proof of Proposition 3.1.3 can be taken to correspond line-by-line to a purely modal derivation in **S4**:

Proposition 3.2.2. For all φ , $\mathbf{S4} \vdash \neg\Box(\varphi \leftrightarrow \neg\Box\varphi)$.

Proof. Reasoning in **S4**:

- i) $\Box(\varphi \leftrightarrow \neg\Box\varphi) \vdash \varphi \leftrightarrow \neg\Box\varphi$ (3.12), **T**
- ii) $\Box(\varphi \leftrightarrow \neg\Box\varphi) \vdash \Box\varphi \rightarrow \varphi$ **T**
- iii) $\Box(\varphi \leftrightarrow \neg\Box\varphi) \vdash \neg\Box\varphi$ i), ii)

¹³Although **SNec** (cf. [104]) is a derivable rule in any normal modal logic containing **S4**, it is not admissible in **T** since it allows us to derive $\Box\varphi \rightarrow \Box\Box\varphi$.

- | | | |
|------|--|-------------------|
| iv) | $\Box(\varphi \leftrightarrow \neg\Box\varphi) \vdash \varphi$ | i), iii) |
| v) | $\Box(\varphi \leftrightarrow \neg\Box\varphi) \vdash \Box\varphi$ | SNec , iv) |
| vi) | $\Box(\varphi \leftrightarrow \neg\Box\varphi) \vdash \perp$ | iii), iv) |
| vii) | $\vdash \neg\Box(\varphi \leftrightarrow \neg\Box\varphi)$ | vi) |

□

The necessity of having to use the **SNec** rule in modal logic to mirror the proof of Proposition 3.1.3 serves to highlight the significance of the outer occurrence of $\Box$ in the reductio hypothesis $\Box(\varphi \leftrightarrow \neg\Box\varphi)$ employed in Proposition 3.2.2. For if this operator were not present, the application of **SNec** at step v) would not be licensed. As discussed in Chapter 2, this can be taken to reflect the fact that it is generally not permissible to conclude that ψ is necessary from hypotheses Δ which are not themselves necessary. The same constraint would appear to apply here if we interpret $\Box$ as expressing knowledge or informal provability – e.g. even if ψ follows logically from Δ , we should not be able to conclude that ψ is informally provable (in a categorical sense) unless this is also true of the hypotheses on which its derivation depends. In the context of the Kowner, the relevant moral can thus be put as follows: since the reasoning leading to a contradiction requires the internalization of reasoning which follows from the existence of a self-

referential statement about knowledge, the proof implicitly requires that this statement be not only true, but also known to be true.

In the context of a theory like T_3 , there is no reason to distinguish between the truth of an arithmetic self-referential statement like (3.12) and its knowability. For since the existence of such statements is a mathematical consequence of the Diagonal Lemma (which is provable in $\mathbf{Q}$), there is no need to derive $K(\ulcorner \delta_3 \leftrightarrow K(\ulcorner \delta_3 \urcorner) \urcorner)$ in the proof of Proposition 3.1.3. Note, however, that the use of the Nec rule at step vi) can be taken to implicitly reflect the fact that this statement is derivable in the system.^{14,15} The con-

¹⁴A similar comment applies to putatively more conservative theories such as T_1 : since, e.g., $\delta_0 \leftrightarrow K(\ulcorner \neg \delta_0 \urcorner)$ is derivable in $\mathbf{Q}$, trivially $T_1 \vdash (K(\ulcorner \top \urcorner) \rightarrow \top) \rightarrow (\delta_0 \leftrightarrow K(\ulcorner \neg \delta_0 \urcorner))$ from which $T_1 \vdash K(\ulcorner \delta_0 \leftrightarrow K(\ulcorner \neg \delta_0 \urcorner) \urcorner)$ follows via U and I. Thus although the use of I at step v) implicitly relies on the fact that $\delta_0 \leftrightarrow K(\ulcorner \neg \delta_0 \urcorner)$ is provable in T_1 , again there is no need to explicitly derive this latter statement. The situation for T_2 is somewhat more complex. Note that since the Diagonal Lemma is provable in $\mathbf{Q}$ we have $\vdash_L Q \rightarrow (\delta \leftrightarrow \neg K(\ulcorner \delta_0 \urcorner))$ from which it follows via Log that $T_3 \vdash K(\ulcorner Q \rightarrow (\delta \leftrightarrow \neg K(\ulcorner \delta_0 \urcorner)) \urcorner)$. But although $\mathbf{Q} \subseteq T_2$, this theory provides no mechanism for internalizing the axioms of $\mathbf{Q}$ themselves. On the one hand, this may be taken to reflect the strength of Proposition 3.1.2 as it shows that in order to get a contradiction, we only have to assume that i can reason hypothetically in $\mathbf{Q}$, not that he believes its axioms. But on the other hand, $\mathbf{Q}$ is already Σ_1 -complete, from which it follows that if $T_2 \vdash \varphi$, then $T_2 \vdash Prov_{T_2}(\ulcorner \varphi \urcorner)$. On this basis, it seems reasonable to expect that i ought to be able to conclude that he knows φ from the fact that he can prove $Prov_{T_2}(\ulcorner \varphi \urcorner)$. Thus there is an important respect in which T_2 seems to be too weak to be a feasible theory of i 's knowledge in that it fails to provide a mechanism for categorically internalizing arithmetic reasoning.

¹⁵A related but more subtle point can also be made about the justification of Nec and the derivability of **SNec** in **S4**. Note in particular that if we might attempt to justify the former rule by an analog to the traditional inductive proof of the admissibility of **SNec** in **S4**, we find that an analog of the 4 axiom is required. For instance, suppose that we start with the putatively more conservative assumption rule Nec^- that $\vdash K(\ulcorner \varphi \urcorner)$ only in the case where φ is an axiom of T_3 . We might then try to argue inductively that if φ follows from some finite set of axioms Γ via a derivation of length n and that we have demonstrated $K(\ulcorner \Gamma \wedge \Delta \urcorner) \vdash \psi \Rightarrow K(\ulcorner \Gamma \wedge \Delta \urcorner) \vdash K(\ulcorner \psi \urcorner)$ for all derivation of length $n - 1$, then we should be able to conclude that $K(\ulcorner \Gamma \wedge \Gamma \urcorner) \vdash K(\ulcorner \varphi \urcorner)$. In order to carry out the induction step,

trast between the arithmetic treatment of self-reference and its status with respect to modal systems like **T** or **S4** be sharpened by noting that whereas T_3 is inconsistent, $\varphi \leftrightarrow \neg\Box\varphi$ is consistent with **T** or **S4** (this can be easily checked by constructing a Kripke model).

With this said, it is reasonable to assume that at the time of [75] the consensus was that self-reference was foreign to propositional modal logic. However, it is now known that there are at least two methods for ensuring that statements such as (3.12) are provable in a system of propositional modal logic. On the one hand, we can introduce the so-called *diagonalization operators* of Smoryński [98]. Such an operator is a function δ_φ which given a modal formula φ containing the propositional letters $Q_1, \dots, Q_n$ and P such that all occurrences of P are within the scope of $\Box$ returns a formula denoted by $\delta_\varphi(P, Q_1, \dots, Q_n)$. If **S** is a normal modal logic, we can define its *diagonal extension* **S**^{*} by adding as axioms all the sentences of the form

$$(3.13) \quad \delta_\varphi(Q_1, \dots, Q_n) \leftrightarrow \varphi(\delta_\varphi(Q_1, \dots, Q_n), P_1, \dots, P_n)$$

we must assume that T_3 satisfies $K(\ulcorner\varphi\urcorner) \rightarrow K(\ulcorner K(\ulcorner\varphi\urcorner)\urcorner)$. Although this principle is not explicitly assumed in any of the forms of the Knower we have considered, the derivability of the schema $Prov_T(\ulcorner\varphi\urcorner) \rightarrow Prov_T(\ulcorner Prov_T(\ulcorner\varphi\urcorner)\urcorner)$ is standardly taken as one of the adequacy on arithmetic proof predicate. But as we have seen, it is not at all clear that such a principle will be derivable in a very weak arithmetic system like **Q**. Hence in adopting *Nec* as a principle governing the behavior of $K(x)$, we are implicitly assuming that $K(x)$ bears a greater resemblance to traditionally provability predicates than is conventionally acknowledged. *Pace* note 14, similar comments apply to the justification of the axiom **I** and thus to T_1 as well.

and closing under modus ponens and **Nec**. If we start out with the modal logic **T**, it's diagonal extension **T*** can be seen to be inconsistent by taking φ to be $\delta_{\neg\Box P}$ and reasoning in the manner of Proposition 3.2.1.

Self-referential statements can also be introduced into modal logic by adopting the familiar logic **GL** – i.e. the normal system containing the axioms **K**, **4** and the Löb axiom

$$(\mathbf{L}) \quad \Box(\Box\varphi \rightarrow \varphi) \rightarrow \Box\varphi$$

The connections between **GL** and arithmetic provability have already been highlighted above. There is also a well known link between **GL** and self-reference in the form of the De Jongh-Sambin Fixed Point Theorem (cf. [7], ch. 8), a simple case of which is as follows:

Theorem 3.2.1. Suppose that φ is formula of $\mathcal{L}_\Box$ such that all instances of the propositional letter P occur within the scope of a modal operator. Then there exists a formula $D_{\varphi(P)}$ such that

$$\mathbf{GL} \vdash D_{\varphi(P)} \leftrightarrow \varphi[D_{\varphi(P)}/P]$$

where $\varphi[D_{\varphi(P)}/P]$ denotes the result of substituting $D_{\varphi(P)}$ for all instances of P in φ .

It can be seen that **GL** is inconsistent with **T** either by applying Theorem 3.2.1 to find a sentence satisfying (3.12) and then reasoning in the manner

of Proposition 3.2.1 or 3.2.2 or by simply taking $\varphi \equiv \perp$ in $\mathbf{L}$ and then using $\mathbf{T}$ and $\mathbf{Nec}$ to derive $\mathbf{U}$.

These developments show that contrary to what Montague suggests in [75], there are ways of developing Knower-like paradoxes in a purely modal setting as long as we assume the availability of a device for generating self-referential sentences of the appropriate form. This suggests that the locus of inconsistency does not lie with any feature particular to the treatment of modality as a predicate of sentences, but rather with the use of an arithmetic theory for securing the existence of self referential statements in conjunction with otherwise plausible epistemic principles like $\mathbf{T}$ and $\mathbf{Nec}$. Further evidence for the fact that a syntactic treatment of modality is orthogonal to issues raised by the Knower is provided by Skyrms's [97] formulation of a consistent metalinguistic treatment of modality wherein a operator satisfying principles similar to those employed in the knower is treated as a predicate of sentences directly through the use of a quotation forming function. It thus seems reasonable to concur with Érgé [25] (p. 37) in criticizing Montague for concluding too hastily that the source of the paradox lies with the syntactic treatment of modalities as opposed to the provability of the Diagonal Lemma in the theory which is employed to formalize reasoning about syntax.

Absent a prior account of how self-referential statements like (3.12) arise out of our ordinary practices, however, there seems to be little motivation for adopting a logic like **GL** or **T*** as a framework in which to reasoning about knowledge. Of course M&K do provide such an account in the case of self-referential statements like (3.2) and (3.4) which appear in the Hangman Paradox – i.e. these sentences arise through the use of demonstrative devices such as “this decree.” But there is an important difference between this process for generating self-referential sentences and the manner in which such sentences are produced in arithmetic theories like T_3 . For in the former case, the self-reference is produced by a sort of linguistic fiat (i.e. by attempting to stipulate that a certain use of a demonstrative refers to a particular utterance), whereas in the latter case, the existence of statements like δ_0 is treated in the proof of Proposition 3.1.1 as a formal consequence of a mathematical theory.

Based on the foregoing modal reconstruction of Proposition 3.1.3, we can now see that the provability (and not just the truth) of such a statement is essential if we wish to argue for the plausibility of the principles on which the Knower is based. For instance, if the statement $\delta_3 \leftrightarrow \neg K(\ulcorner \delta_3 \urcorner)$ were only a contingent truth which we had subsumed into T_3 , then the case for adopting the Nec rule as a principle about knowledge would be substantially

weakened. For note that i might fail to know such a statement either in virtue of failing to have observed the relevant act of demonstration by which δ_3 was introduced or by failing to possess irrefutable evidence that this act succeeded in referring to the relevant sentence.¹⁶ As such, there is no reason to presume that i can come to *know* (as opposed to, say, rationally believe) statements which follow from such self-referential statements unless these statements are accorded the status of mathematical theorems. It hence follows that if the self-referential statement employed in the derivation of the Knower lacks this property, then it is not clear that any of the forms of the Knower we have considered demonstrate a genuine clash among obviously valid epistemic principles.

It thus appears that within the epistemic tradition, it is perfectly reasonable to respond to the Knower by denying that there *provably* exist self-referential statements involving the notion which $K(x)$ is intended to express. From a rhetorical standpoint, this corresponds to the response M&K adopt with respect to the Hangman in the sense that they take the inconsistency of (3.3) with T, I and U to show that it is impossible to utter a decree in natural

¹⁶Note that what is at issue is not whether the use of a given usage of a demonstrative like “this decree” or “this sentence” succeeds in referring to a particular syntactic token, but rather whether an agent can rationally doubt whether such a usage succeeds in referring. In order to see that there are grounds for such doubt, we need look no further than the fact that I might very well attempt to use an expression like “this rabbit” to refer to an object which I had hallucinated.

language which successfully expresses (3.2). But as noted above, most commentators have not taken this path in response to the Knower, but rather have looked for reasons to reject one or more of the epistemic principles on which the derivations given in Section 1 are based. I will now turn to the task of reconstructing the derivation of Proposition 3.1.3 in QLP so as to deepen our understanding of what these options entail. On the assumption that we have not already chosen to resolve the Knower by denying the provability of statements like (3.12), what will emerge is a strengthened argument in favor of the consensus view that U ought to be rejected as a principle of everyday epistemology.

3.3 The reconstruction of the Knower in QLP

Like propositional modal logic, QLP also does not provide an obvious mechanism for reasoning about its own syntax. If we wish to reconstruct the reasoning of the Knower, it is thus reasonable to proceed in the same manner as Section 2 – i.e. by starting out with the assumption that an appropriate certain self-referential statement *is* provable and then attempting to reason in the same manner as Proposition 3.2.2. Since this is a derivation in S4, we are aided in this regard by Proposition 3.13 of Chapter 2 which I reproduce here for reference:

Proposition 3.3.1. Define the mapping $(\cdot)^\exists$ between the language of **S4** and **QLP** which is the identity map on atomic statements, commutes with propositional quantifiers and is such that $(\Box\varphi)^\exists = (\exists x)x : \varphi^\exists$. Then **S4** $\vdash \varphi$ if and only if **QLP**($\mathcal{P}$) $\vdash \varphi^\exists$ for any axiomatically appropriate primitive term specification $\mathcal{P}$.

From this it follows directly from Proposition 3.2.2 that **QLP** $\vdash \neg(\exists y)y : [\varphi \leftrightarrow \neg(\exists x)x : \varphi]$. But in addition, it can be readily seen that this result follows via a proof whose structure more closely matches that of this proposition.¹⁷

Proposition 3.3.2. For all φ , **QLP** $\vdash \neg(\exists y)y : [\varphi \leftrightarrow \neg(\exists x)x : \varphi]$.

Proof. Reasoning in **QLP**:

- i) $y : (\varphi \leftrightarrow \neg(\exists x)x : \varphi) \vdash \varphi \leftrightarrow \neg(\exists x)x : \varphi$ LP3
- ii) $y : (\varphi \leftrightarrow \neg(\exists x)x : \varphi) \vdash (\exists x)x : \varphi \rightarrow \varphi$ derivable in **QLP**
- iii) $y : (\varphi \leftrightarrow \neg(\exists x)x : \varphi) \vdash \neg(\exists x)x : \varphi$ ii), iii)
- iv) $y : (\varphi \leftrightarrow \neg(\exists x)x : \varphi) \vdash \varphi$ i), iv)
- v) $y : (\varphi \leftrightarrow \neg(\exists x)x : \varphi) \vdash t(y) : \varphi$ for some $t(y)$ via the Lifting Lemma
- v') $y : (\varphi \leftrightarrow \neg(\exists x)x : \varphi) \vdash (\exists x)x : \varphi$ QLP3
- vi) $y : (\varphi \leftrightarrow \neg(\exists x)x : \varphi) \vdash \perp$ iv), vi')
- vii) $\vdash \neg y : (\varphi \leftrightarrow \neg(\exists x)x : \varphi)$ vi)

vii') $\vdash \neg(\exists y)y : [\varphi \leftrightarrow \neg(\exists x)x : \varphi]$ RQLP2

□

With the exception of the quantifier steps v') and v''), this derivation also shares the structure of Proposition 3.1.3. As presented, however, the argument is an enthymeme. In particular, neither steps ii) nor v) correspond to QLP axioms nor are they derivable from the preceding steps by QLP rules. The first of these gaps may be filled by the following additional derivation:

- | | | |
|----|--|--------|
| 1) | $\vdash x : \varphi \rightarrow \varphi$ | LP3 |
| 2) | $\vdash (\forall x)(x : \varphi \rightarrow \varphi)$ | UG |
| 3) | $\vdash (\forall x)(x : \varphi \rightarrow \varphi) \rightarrow ((\exists x)x : \varphi \rightarrow \varphi)$ | QLP4 |
| 4) | $\vdash (\exists x)x : \varphi \rightarrow \varphi$ | 3), 4) |

Recall that at step v) in the original derivation of Proposition 3.1.3, we are able to conclude $K(\ulcorner \delta_3 \urcorner)$ from the proceeding proof of δ_3 via the rule Nec. Although QLP does not have a single rule which allows for internalization of this form, it does satisfy the Lifting Lemma, meaning that whenever a statement ψ is derivable, we can construct an appropriate proof term s such

¹⁷It is also possible to obtain a similar reconstruction of T-derivation corresponding to the proof of Proposition 3.2.1. I have elected to base the presentation on Proposition 3.2.2 to more directly exhibit the relationship between the S4 necessitation rule and the Lifting Lemma of QLP.

that $s : \psi$ is derivable. It is virtue this that we know that the term $t(y)$ introduced at step v) of the current derivation exists. In order to construct this term explicitly, we must internalize the reasoning of the steps i)-v). Steps 1), 3) and 4) involve only propositional reasoning and are thus straightforward to internalize. However, in order to internalize step ii), we must show that the derivation just given involving the application of the QLP UG rule can be internalized. This can be accomplished using the rule GJUG which was introduced in Chapter 2 and which I reproduce here for reference

$$\begin{aligned} \text{(GJUG)} \quad \vec{s} : \vec{\Gamma} \vdash t(x) : \varphi(x) \quad \therefore \quad \vec{s} : \vec{\Gamma} \vdash \langle t(x)\forall x \rangle : (\forall x)\varphi(x) \text{ where} \\ x \notin \text{FV}(\vec{s} : \vec{\psi}) \end{aligned}$$

The requisite derivation using this rule is as follows:

- | | | |
|----|--|-------------|
| 5) | $\vdash x : \varphi \rightarrow \varphi$ | LP3 |
| 6) | $\vdash r(x) : (x : \varphi \rightarrow \varphi)$ | RQLP3 |
| 7) | $\vdash \langle r(x)\forall x \rangle : (\forall x)(x : \varphi \rightarrow \varphi)$ | GJUG, 6) |
| 8) | $\vdash q : (\forall x)(x : \varphi \rightarrow \varphi) \rightarrow ((\exists x)x : \varphi \rightarrow \varphi)$ | RQLP3 |
| 9) | $\vdash q \cdot \langle r(x)\forall x \rangle : ((\exists x)x : \varphi \rightarrow \varphi)$ | LP2, 7), 8) |

It is now routine to establish that the term $t(y)$ appearing at step v) of the original derivation can be taken to be of the form $(a_1 \cdot y) \cdot ((b \cdot (q \cdot (r(x)\forall x))) \cdot (a_2 \cdot y))$ where $r(x)$ and q are primitive proof terms which justify

the formulas displayed in the derivation and a_1, a_2 and b are primitive proof terms for propositional tautologies.¹⁸

What is of more immediate interest is the status of the two foregoing sub-derivations and the formulas which they demonstrate. Step 4) corresponds to the principle

$$(\mathbf{T}_q) \ (\exists x)x : \varphi \rightarrow \varphi$$

mentioned in Section 1 which is the image of the modal axiom $\mathbf{T}$ under the mapping $(\cdot)^\exists$. Similarly, from the formula on line 9) we may obtain the following by QLP2 (i.e. existential generalization) in QLP:

$$(\mathbf{U}_q) \ (\exists y)y : ((\exists x)x : \varphi \rightarrow \varphi)$$

This corresponds to the image under $(\cdot)^\exists$ of the modal principle $\mathbf{U}$ considered in the prior section. We have seen these formulas both play a prominent role in the various forms of the Knower. But in QLP, neither $\mathbf{T}_q$ nor $\mathbf{U}_q$ is assumed as an axiom. They must rather be derived in the manner just exhibited.

On the standard interpretation of QLP quantifiers, $\mathbf{T}_q$ asserts what I have called the factivity of justification holds universally – i.e. if for all justifications x , if x verifies φ , then φ is true. As discussed in Chapter 1, this former

¹⁸Respectively: $a_1 : (A \leftrightarrow \neg B) \rightarrow (B \rightarrow \neg A)$, $a_2 : (A \leftrightarrow \neg B) \rightarrow (A \rightarrow \neg B)$ and $b : (B \rightarrow A) \rightarrow ((B \rightarrow \neg A) \rightarrow \neg B)$.

property appears to be central to our understanding of justification understood in the sense of informal proof. I have argued above that the setting of the Knower Paradox requires us to view $K(x)$ as expressing existential quantification over some sort of justification-like objects such as proofs in T_3 . But although we have seen that there are reasons to hold that factivity is not an intrinsic feature of formal proof, it is notable that nothing about formulation of the Knower as a paradox about the everyday conception of knowledge predisposes us to think of justification as not being factive. In fact, the acceptance of T as a principle about knowledge seems to codify exactly this feature of knowledge. As such, it seems that there is a reasonable basis for accepting $\mathbf{T}_q$ in the setting of the Knower.

Assessing the status of $\mathbf{U}_q$ in this context is considerably more delicate. This principle can be taken to assert that for each φ , there is a justification which verifies that all justifications of φ are factive. In order to make an intuitive case in favor of such a principle appears to require making sense of what it means to provide a justification for a statement which quantifies over all justifications. The fundamental difficulty involved can be seen by asking whether we can construct an informally rigorous argument for $\mathbf{U}_q$ by reflecting on the proof 1) - 4) of $\mathbf{T}_q$ given above. The crux of this argument is the application of the UG rule for proof variables applied at step 2). In Chapter

I argued for the soundness of this rule based our intuitive acceptance of the universal closure of the LP application and proof checker axioms. However, in accepting the soundness of the rule applied to $x : \varphi \rightarrow \varphi$ where x is a free proof variable, we are committing ourselves to the factivity of all the objects which fall under our background conception of justification.

As I will discuss further below, however, it seems likely that we do not possess a completely determinate conception of what it is to be such an object. In particular, although the setting of the Knower encourages us to view formal proofs in a system such as T_3 as justifications, our general notion of justification or proof is certainly wider than this. For not only are there many statements which are not statable in the language of arithmetic of which we think we are justified in believing, but it is also possible that we might possess justification for arithmetic statements which are not provable in T_3 . Upon noting this, it seems that we might reasonably hesitate in accepting $\mathbf{U}_q$. For if our informal concept of justification is indeed indeterminate or open-ended in the manner just indicated, then we might reasonably doubt whether *all* of the evidentiary considerations which we might take to fall under this concept are indeed factive.

These observations may prove to be significant to the Knower since each of the derivations we have seen previously require either the principle $\mathbf{U}$ or

rule Nec which, in conjunction with T, is sufficient to derive U. It is notable, however, that in the course of the QLP derivation of U_q , the rule GJUG appears to be necessary in order to internalize the use of UG at line 2) above. As we will see in the next section, in the case where φ corresponds to a propositional letter (and is thus not derivable in QLP itself), this is in fact the case. This result demonstrates that without GJUG, QLP is not capable of deriving the relevant instances of U_q to reconstruct the reasoning of the Knower. Thus if we already think that there may be a rationale for rejecting this principle, it is reasonable to consider further the significance of GJUG itself. For although such a principle has not traditionally been flagged as one of the premises on which the paradox depends, it appears to be necessary in the context of any proposal which seeks to reconstruct the reasoning of the Knower on the assumption that $K(x)$ is intended to existential quantification over justifications.

3.4 Arithmetic Semantics for QLP

I suggested above that the setting of the Knower is apparently neutral with respect to whether epistemic principles like T and U ought to be understood in terms of informal or formal proofs. In order to investigate the status of the various principles on which the various derivations of the Knower Para-

dox depend, it will also be useful to consider which axioms of QLP remain valid when interpreted as statements about formal proofs in an arithmetic theory T . One way to do this is to formulate an arithmetic interpretation $(\cdot)^\circ$ of the language of QLP which interprets quantifiers in this language as ranging over (codes of) arithmetic proofs and which maps explicit modalities of the form $t : \varphi$ to arithmetic formulas of the form $Proof_T(t^\circ, \varphi^\circ)$. What we will see in this section is that for an appropriate formulation of $(\cdot)^\circ$ and arithmetic theory T , all of the axioms/rules of QLP will be true/sound with respect to the standard model of arithmetic except for GJUG. As one might expect, this fact is closely related to the Second Incompleteness Theorem, or more directly, to the internalized form of Löb's Theorem. For this reason, it will be simplest to take $T = \text{PA}$ so that we may assume that these results are provable. However, I will suggest in the next section that the formulation of an arithmetic interpretation for QLP casts some additional light on the acceptability of GJUG independently of the precise choice of T .

In specifying such an arithmetic interpretation for QLP we face several problems. First, while quantifiers in the language of QLP are intended to range over justifications or proofs, the quantifiers in the language of arithmetic range over natural numbers. And, second, we already have some *a priori* reasons to suspect that it is not possible to provide a natural arith-

arithmetic interpretation such that the image of *all* of the theorems of QLP are arithmetically provable (or even true in the standard model) at once. For as we have seen, both $\mathbf{T}_q$ and $\mathbf{U}_q$ are theorems of QLP. The pre-images of these statements under the embedding $(\cdot)^\exists$ respectively corresponds to $\Box\varphi \rightarrow \varphi$ and $\Box(\Box\varphi \rightarrow \varphi)$. These are the modal analogues of the principles T and U which we have seen in Theorem 3.1.3 lead to a contradiction when added to a system to which the Diagonal Lemma applies.

In seeking to formulate an arithmetic semantics for QLP we thus should not be led by a desire to demonstrate that the interpretation of all theorems of QLP are either true or provable for all choices of $(\cdot)^\circ$. Rather, we should attempt to define $(\cdot)^\circ$ so that, to as great an extent as possible, the meaning of QLP statements are preserved when the relation $t : \varphi$ is understood as expressing the relation which holds between a formal proof and its conclusion. To this end, I will attempt to base the definition of $(\cdot)^\circ$ as closely as possible on the definition of an arithmetic interpretation of LP given in Chapter 2. We will see below that it is possible to define an interpretation so that if $\text{QLP}_0(\mathcal{P}) \vdash \varphi$, then $(\varphi)^\circ$ is true in standard model for all $(\cdot)^\circ$ which satisfy the primitive term specification $\mathcal{P}$. However, we will also see that relative to this definition, not only is GJUG not a sound rule of inference, but that adding the images of all theorems of QLP to PA results in an inconsistent

theory.

Let $\text{Prf}(x, y)$ be a proof predicate for PA which satisfies the same conditions assumed in Section 2.2.4 – i.e.

- $\text{Prf}(x, y)$ is provably Δ_1 in PA
- for every arithmetic formula φ , $\text{PA} \vdash \varphi \iff$ for some $n \in \mathbb{N}$, $N \models \text{Prf}(\bar{n}, \overline{\ulcorner \varphi \urcorner})$
- If we let $T(n) = \{k \mid N \models \text{Prf}(\bar{n}, \bar{k})\}$, then $\text{Prf}(x, y)$ satisfies
 - for all n, m there is a k such that $T(n) \cup T(m) \subseteq T(k)$
 - $T(n)$ is finite for all n .

It also will be useful to assume that the arithmetic language in which we are working contains terms $\mathbf{f}_0, \mathbf{f}_1, \dots$ which represent all primitive recursive functions. I will accordingly assume that PA contains the defining axioms for all these functions (see [98] for details of how this can be handled). It follows from the fact that $\text{Prf}(x, y)$ is provably Δ_1 that the set of $P = \{n \mid N \models \exists y \text{Prf}(\bar{n}, y)\}$ is recursive. There is hence a computable function $p : \mathbb{N} \rightarrow \mathbb{N}$ which enumerates P bijectively – i.e. $P = \{p(n) \mid n \in \mathbb{N}\}$ and if $p(n) = p(m)$, then $n = m$. Let $\mathbf{p}$ be a name for a $\mathcal{L}_a$ -representation of this function. We assume that the official names for proof variables in $\mathcal{L}_{\text{QLP-}}$ are $x_0, x_1, \dots$ and

the official names for arithmetic variables in $\mathcal{L}_a$ are $y_0, y_1, \dots$. We then set $(x_i)^\circ = \mathbf{p}(y_i)$. In the case where x_i occurs free in a QLP formula $\varphi(x_i)$, this will mean that the y_i will occur free in $(\varphi(x_i))^\circ$ inside a term of the form $\mathbf{p}(y_i)$.

Suppose $\mathcal{P}$ is an axiomatically appropriate primitive term specification. Thus for each instance of a QLP axiom φ , there is thus some primitive proof term a such that $\varphi \in \mathcal{P}(a)$ and $\text{FV}(a) = \text{FV}(\varphi)$. We wish the images of statements of the form $a(x_{k_1}, \dots, x_{k_n}) : \varphi(x_{k_1}, \dots, x_{k_n})$ such that $\varphi(x_{k_1}, \dots, x_{k_n}) \in \mathcal{P}(a(x_{k_1}, \dots, x_{k_n}))$ to get mapped to arithmetic open sentences which are true for all numerical substitution instances of their free variables. This means that the image of a primitive proof term must itself be a function mapping numbers into codes of proofs. For instance, if $\varphi(x_{k_1}, \dots, x_{k_n})$ is an axiom of QLP and we have defined $(\cdot)^\circ$ so that $\varphi(x_{k_1}, \dots, x_{k_n})^\circ$ is an open sentence of $\mathcal{L}_a$ provable in PA, then the interpretation of $a(x_{k_1}, \dots, x_{k_n})$ should be a function $a_{\varphi(x_{k_1}, \dots, x_{k_n})}(x_{k_1}^\circ, \dots, x_{k_n}^\circ)$ which for every assignment of QLP terms to $x_{k_1}, \dots, x_{k_n}$, returns the code of an arithmetic proof of the image of the corresponding substitution instance of $\varphi(x_{k_1}, \dots, x_{k_n})$.

We must also provide a means of interpreting the proof operations $\cdot, +, !$ and $\langle \cdot \forall \cdot \rangle$ of $\mathcal{L}_{\text{QLP}}$. In the case of the first three operators, this can be accom-

plished in a manner similar to Artemov's original arithmetic interpretation of LP via the use of three functions $m(y_1, y_2)$, $s(y_1, y_2)$ and $c(y)$. In particular, $m(y_1, y_2)$ returns the least godel-number of a proof containing all instances of ψ such that $\ulcorner \varphi \rightarrow \psi \urcorner \in T(y_1)$ and $\ulcorner \varphi \urcorner \in T(y_2)$. $s(y_1, y_2)$ returns the least godel-number of a proof containing $T(y_1)$ and $T(y_2)$ and $c(y)$ returns the least godel-number of a proof z such that $\text{Prf}(z, \overline{\ulcorner \varphi \urcorner}) \in T(y)$ for all $\ulcorner \varphi \urcorner \in T(y_1)$. Finally, in order to interpret the universal verifier symbol $\langle \cdot \forall \cdot \rangle$, we define a function $g(y_1, y_2)$ which, if y_1 is the godel number of a proof and y_2 is the godel number of a variable y , returns the least godel number of a proof such z such that $\text{Prf}(z, \overline{\ulcorner (\forall y) \varphi \urcorner}) \in T(z)$ for all $\varphi \in T(y_1)$ such that $y \in \text{FV}(\varphi)$. Following [3], it is routine to show that these functions are primitive recursive and may hence be represented by arithmetic terms $\mathbf{m}$, $\mathbf{s}$, $\mathbf{c}$ and $\mathbf{g}$.

Following Boolos [7], let $s(x, y, z)$ be the primitive recursive function $su(x, y, z)$ such that for all i, j, k $su(i, j, k)$ is the godel number of the result of substituting the numeral $\bar{i}$ for the j th free variable in formula with godel number k . Putting these pieces together we can now specify what it is to be an arithmetic interpretation for QLP and how an interpretation applies to proof terms and formulas.

Definition 3.4.1. An arithmetic interpretation for QLP is a mapping $(\cdot)^\circ$ defined on both proof terms and formulas, defined relative to the definitions

of $\text{Prf}(y_1, y_2)$, $m(y_1, y_2)$, $p(y_1, y_2)$, $c(y)$ and $g(y_1, y_2)$ just given. The value of t° define inductively as follows:

- $x_i^\circ = \mathbf{p}(y_i)$
- $(a(x_1, \dots, x_n))^\circ = \mathbf{f}(x_1^\circ, \dots, x_n^\circ)$ where the latter is a name for an arbitrary arithmetically representable n -ary function
- $(t \cdot s)^\circ = \mathbf{m}(t^\circ, s^\circ)$, $(t + s)^\circ = \mathbf{s}(t^\circ, s^\circ)$, $(!t)^\circ = \mathbf{c}(t^\circ)$, $(\langle t \forall x_i \rangle)^\circ = \mathbf{g}(\overline{\text{Prf}(t^\circ)}, \overline{\text{Prf}(y_i)})^{19}$

The value of φ° is defined inductively as follows:

- $(P_i)^\circ$ is an arbitrary closed sentence of $\mathcal{L}_a$
- $(\varphi \rightarrow \psi)^\circ = \varphi^\circ \rightarrow \psi^\circ$, $(\neg \varphi)^\circ = \neg \varphi^\circ$
- $((\forall x_i) \varphi)^\circ = (\forall y_i) \varphi^\circ$, $((\exists x_i) \varphi)^\circ = (\exists y_i) \varphi^\circ$
- $t : \varphi(x_{k_1}, \dots, x_{k_m}) =$
 $\text{Prf}(t^\circ, \mathbf{su}(x_{k_m}^\circ, \overline{k_m}, \dots, \mathbf{su}(x_2^\circ, \overline{k_2}, \mathbf{su}(x_1^\circ, \overline{k_1}, \overline{\varphi(x_{k_1}^\circ, \dots, x_{k_m}^\circ)})) \dots))$

¹⁹Note that unlike the other other functions used to interpret proof terms $\mathbf{g}$ is defined so that it applies not to the result of interpreting its arguments relative to $(\cdot)^\circ$, but rather their godel numbers. This reflects the fact that occurrences of x in the uniform verifier term $\langle t \forall x \rangle$ are treated as bound. This manner of treating $\langle \cdot \forall \cdot \rangle$ is in some sense arbitrary because we will see below that it does not lead to an arithmetically valid interpretation of GJUG. However, we see below that this is inevitable because GJUG can be used mediate an arithmetically unsound inference between statements not involving the uniform verifier symbol.

It will be useful to observe the following differences between the way proof variables are handled with respect to arithmetic interpretations for LP versus for QLP. Recall that an arithmetic interpretation $(\cdot)^*$ for LP associates each proof variable x with a fixed natural number, meaning that the image of an open LP formula under $(\cdot)^*$ will be a closed arithmetic formula. As is the case for LP, there are instances of QLP axioms containing free proof variables such as

$$(3.14) \quad x : P \rightarrow P$$

But under an arithmetic interpretation $(\cdot)^\circ$ for QLP, this formula will get mapped to an arithmetic open sentence of the form

$$(3.15) \quad \text{Prf}(\mathbf{p}(y), \overline{\lceil P^\circ \rceil}) \rightarrow P^\circ$$

Since the former formula will be valid with respect to the relational semantics for QLP given in Chapter 2, we would also like it to come out as arithmetically valid. The most straightforward way of achieving this is to view free arithmetical variables as universally bound in the standard manner. In particular, I will say that a QLP sentence $\varphi(x_{k_1}, \dots, x_{k_m})$ with free variables displayed is *arithmetically valid with respect to $(\cdot)^\circ$* if $N \models_{v[y_{k_1}/n_1, \dots, y_{k_m}/n_m]} (\varphi)^\circ(y_{k_1}, \dots, y_{k_m})$ for all $n_1, \dots, n_m \in \mathbb{N}$. In case φ° is arithmetically valid with respect to all interpretations $(\cdot)^\circ$, I will say φ° is simply *arithmetically*

valid. If $\mathcal{P}$ is a primitive term specification, we similarly say that an interpretation $(\cdot)^\circ$ is $\mathcal{P}$ -arithmetically valid if $a : \varphi$ is arithmetically valid with respect to $\mathcal{P}$ for all a, φ such that $\varphi \in \mathcal{P}(a)$.

On the route to demonstrating the arithmetic soundness of QLP_0 , we must first convince ourselves that if $\mathcal{P}$ is axiomatically appropriate, then there exist $\mathcal{P}$ -arithmetically valid interpretations. As described above, this essentially involves showing that if the formula φ with free variables $x_{k_1}, \dots, x_{k_m}$ is an axiom of QLP_0 , and $\varphi \in \mathcal{P}(a)$, then there exists a primitive recursive function a_φ with arguments $y_{k_1}, \dots, y_{k_m}$ such that for all assignments of natural numbers $n_1, \dots, n_m$, $a_\varphi(n_1, \dots, n_m)$ returns a proof of the interpretation of φ when its free variables are mapped to $p(n_1), \dots, p(n_m)$. If we can show that such functions exist for all primitive proof terms a and formulas φ such that $\varphi \in \mathcal{P}(a)$, then we will have obtained a $\mathcal{P}$ -arithmetically valid interpretation.

Since it is routine but tedious to show that such functions exist I will make the case by considering several examples. Consider first the case of the LP reflection axiom LP3 given in (3.14) and recall that its arithmetic interpretation is given by (3.15). Although this is an open arithmetic sentence, it is easy to see that every for all $n \in \mathbb{N}$

$$(3.16) \quad \text{PA} \vdash \text{Prf}(\mathbf{p}(\bar{n}), \ulcorner P^\circ \urcorner) \rightarrow P^\circ$$

For note that we have either i) $N \models \text{Prf}(\mathbf{p}(\bar{n}), \ulcorner P^\circ \urcorner)$ or ii) $N \not\models \text{Prf}(\mathbf{p}(\bar{n}), \ulcorner P^\circ \urcorner)$.

In the first case, $\text{Prf}(\mathbf{p}(\bar{n}), \ulcorner P^\circ \urcorner)$ is a true Δ_1 -sentence and hence $\mathbf{p}(\bar{n})^N$ is indeed the gödel number of a proof of P° in $\mathbf{PA}$. Hence by the Σ_1 -completeness of $\mathbf{PA} \vdash P^\circ$, meaning that (3.16) holds. In the second case, $\text{Prf}(\mathbf{p}(\bar{n}), \ulcorner P^\circ \urcorner)$ is a false Δ_1 sentence, meaning that $\neg \text{Prf}(\mathbf{p}(\bar{n}), \ulcorner P^\circ \urcorner)$ is again a true Σ_1 -sentence. Hence $\mathbf{PA} \vdash \neg \text{Prf}(\mathbf{p}(\bar{n}), \ulcorner P^\circ \urcorner)$ by Σ_1 -completeness and (3.16) again holds.

The foregoing reasoning is uniform in n in the sense that each substitution instance of (3.15) is true. But also note that given $n \in \mathbb{N}$, we can calculate the gödel number of a proof of $\text{Prf}(\mathbf{p}(\bar{n}), \ulcorner P^\circ \urcorner) \rightarrow P^\circ$ via the following procedure: 1) given n , compute $p(n)$ and check whether $\text{Prf}(\mathbf{p}(\bar{n}), \ulcorner P^\circ \urcorner)$ is true (this can be decided because $\text{Prf}(\mathbf{p}(\bar{n}), \ulcorner P^\circ \urcorner)$ is Δ_1); 2) if so, i) $p(n)$ really is the gödel of a proof of P° in $\mathbf{PA}$ so ii) we can construct the gödel number of a proof of this instance of (3.15) by composing the proof denoted by $p(n)$ with a proof of the appropriate instance of the tautology $B \rightarrow (A \rightarrow B)$; 3) if not, then since $p(n)$ does not denote a proof of P° , we can i) obtain the gödel number q of proof of $\neg \text{Prf}(\mathbf{p}(\bar{n}), \ulcorner P^\circ \urcorner)$ by simply examining the structure of the proof denoted by $p(n)$ and noting that it does not have $\ulcorner P^\circ \urcorner$ as a conclusion and we can then ii) construct the gödel of a proof of this instance of (3.15) by composing q with a proof of the appropriate instance of the tautology

$\neg A \rightarrow (A \rightarrow B)$. Since the procedure just described is clearly effective and involves only bounded search, we know there is a primitive recursive function which for every n , returns the godel number of a proof demonstrating (3.16) in PA. It is this function which we take as the axiom function $a_{x:P \rightarrow P}(y)$.

By way of another example, consider the following instance of the axiom QLP1

$$(3.17) \quad (\forall x)\psi(x) \rightarrow \psi(t)$$

where, for simplicity, we assume that ψ contains no free variables other than x . The arithmetic interpretation of this formula is

$$(3.18) \quad (\forall y)\psi^\circ(\mathbf{p}(y)) \rightarrow \psi^\circ(t^\circ)$$

Reasoning in PA, suppose that $(\forall y)\psi^\circ(\mathbf{p}(y))$ is true. This can be seen to be equivalent to a statement of the form $(\forall y)(\text{Pf}(y) \rightarrow \psi^\circ(y))$ where $\text{Pf}(x)$ abbreviates $\exists y' \text{Prf}(y, y')$. Since we are assuming that t° corresponds to an $\mathcal{L}_a$ term, it follows by universal instantiation that $\text{Pf}(t^\circ) \rightarrow \psi^\circ(t^\circ)$. Assuming appropriate definitions of the functions used to define t° , we may show by induction on the structure of t that $\text{Pf}(t^\circ)$ holds. And from this we obtain $\psi^\circ(t^\circ)$. Since in this case the axiom under consideration has no free variables, we may take $a_{(\forall x)\psi(x) \rightarrow \psi(t)}$ to be the constant function which returns a godel number which encodes the proof we have just described.

With these preliminaries out of the way, we may state the arithmetic soundness theorem for QLP_0 .

Theorem 3.4.1. (Arithmetic soundness of QLP_0) Suppose that $\text{QLP}_0(\mathcal{P}) \vdash \varphi$. Then $N \models \varphi^\circ$ for all $\mathcal{P}$ -valid interpretations $(\cdot)^\circ$.

Proof. Let $\mathcal{P}$ be an axiomatically appropriate primitive term specification and $(\cdot)^\circ$ be $\mathcal{P}$ -arithmetically valid. Note that the foregoing discussion demonstrates that for all instance of QLP axioms, whenever $\varphi \in \mathcal{P}(a)$, $N \models \text{Prf}(\overline{a_\varphi}, \overline{\varphi^\circ})$. From this it follows that $\text{PA} \vdash \varphi^\circ$ and hence by the Σ_1 -soundness of PA, that $N \models \varphi^\circ$.²⁰

Turning now to the rules of QLP_0 , the soundness of RQLP3 (i.e. axiom necessitation) follows immediately from the fact that $\mathcal{P}$ is arithmetically valid. The soundness of RLP2 (i.e. modus ponens) also follows immediately from the soundness of modus ponens in N . In order to see that RQLP2 (i.e. Universal Generalization) is an arithmetically sound rule, suppose that $\text{QLP}_0 \vdash \varphi(x)$ (for simplicity, assume that $\text{FV}(\varphi) = \{x\}$). By the induction hypothesis, it follows that $N \models \varphi^\circ(\mathbf{p}(y))$ which, as above, is seen to be equiv-

²⁰Note that we can also argue for this claim in a manner that does not rely on the Σ_1 -soundness of PA. In the case of axioms LP1-LP5, these arguments are similar to those those for the soundness of the LP axioms given by Artemov in [3] with the additional complication that free proof variables must now be interpreted as universally bound. In the case of QLP1-QLP4, the proofs follow the format of the argument given for the provability of (3.17) above, by simply citing the arithmetic validity of the corresponding Hilbert axioms for classical predicate calculus.

alent to $N \models \text{Pf}(y) \rightarrow \varphi^\circ(y)$. By the definition of arithmetic validity, this is equivalent to $N \models_{v[y/n]} \text{Pf}(y) \rightarrow \varphi^\circ(y)$ for all $n \in \mathbb{N}$ where v is an arbitrary arithmetic variable assignment. It hence follows that $N \models (\forall y)\varphi^\circ(\mathbf{p}(y))$. $\square$

Theorem 3.4.1 is essentially a reflection of the fact that the theorems of LP remain arithmetically valid when we interpret their free variables as universally bound free variables over (codes of) arithmetic proofs. We will see in a moment, however, that it already has several implications for the reconstruction of the Knower Paradox in QLP. First, however, it will be useful to note one respect in which QLP_0 and LP contrast sharply with respect to arithmetic interpretability. Recall in particular, that if $(\cdot)^*$ is an LP interpretation, then not only are the images of all theorems of LP true in N , but they are also provable in PA. However, it is easy to see that there is no hope of extending Theorem 3.4.1 to an analogous result for QLP_0 where “true in N ” is replaced with “provable in PA” in the definition of arithmetic validity.

In order to see this, consider the case of the LP reflection axiom with $\varphi \equiv \perp$ – i.e.

$$(3.19) \quad x : \perp \rightarrow \perp \equiv \neg x : \perp$$

Via Universal Generalization in QLP_0 we have that

$$(3.20) \quad \text{QLP}_0 \vdash (\forall x)\neg x : \perp$$

The image of this statement under $(\cdot)^\circ$ is

$$(3.21) \quad (\forall y) \neg \text{Prf}(\mathbf{p}(y), \ulcorner \perp \urcorner)$$

which can be seen to express “for all y , $p(y)$ is not a proof of a contradiction in PA.” Assuming that PA is consistent, this statement is true in the standard model. However, since p enumerates the class P of all arithmetic proofs, (3.21) expresses the consistency of PA and is thus not provable in PA by the Second Incompleteness Theorem (again assuming that PA is itself consistent). It thus follows that although the interpretation (in the QLP sense) of all instances of the LP reflection axiom are true in N , an instance with a free variable in its antecedent like (3.15) is already not provable in PA.

To be more precise about this, it will be useful to record the following forms of familiar results about PA²¹:

Proposition 3.4.1 (Gödel’s Second Incompleteness Theorem). If PA is consistent, then

$$\text{PA} \not\vdash (\forall y) \neg \text{Prf}(\mathbf{p}(y), \ulcorner \perp \urcorner).$$

Proposition 3.4.2 (Internalized Löb’s Theorem). For all $\mathcal{L}_a$ sentences φ ,

$$\text{PA} \vdash (\exists y) \text{Prf}(\mathbf{p}(y), \ulcorner (\exists y) \text{Prf}(\mathbf{p}(y), \ulcorner \varphi \urcorner) \rightarrow \varphi \urcorner) \rightarrow (\exists y) \text{Prf}(\mathbf{p}(y), \ulcorner \varphi \urcorner)$$

On the basis of Proposition 3.4.1, we can now see that although (3.21) is the image of a theorem (3.20), of QLP_0 it is provable in PA only if PA is inconsistent. A number of related results follow in a similar manner.

Proposition 3.4.3. Let $\mathcal{P}$ be axiomatically appropriate and $(\cdot)^\circ$ a $\mathcal{P}$ -arithmetically valid interpretation. Then there are sentences φ such that $\text{QLP}(\mathcal{P}) \vdash \varphi$, but $N \not\models \varphi^\circ$.

Proof. Let $\mathcal{P}$ be as in the hypotheses. Then as above,

$$(3.22) \quad \text{QLP}(\mathcal{P}) \vdash \neg r(x) : (x : \perp)$$

It then follows by GJUG that

$$(3.23) \quad \text{QLP}(\mathcal{P}) \vdash \langle r(x) \forall x \rangle : (\forall x) \neg x : \perp$$

and then by QLP2 (i.e. existential generalization) that

$$(3.24) \quad \text{QLP}(\mathcal{P}) \vdash (\exists y)y : (\forall x) \neg x : \perp$$

Note, however, that the image of (3.24) under $(\cdot)^\circ$ is equivalent to

$$(3.25) \quad (\exists y)\text{Prf}(\mathbf{p}(y), \ulcorner (\exists y)\text{Prf}(\mathbf{p}(y), \ulcorner \perp \urcorner) \urcorner \rightarrow \perp)$$

²¹These formulations follow easy from the traditional proofs (cf., e.g., [8]) as long as we assume that $\mathbf{p}$ is defined in such a way that the fact that, provably in PA , it enumerates all members of the set of P of godel number of arithmetic proofs relative to the definition of $\text{Prf}(y_1, y_2)$.

Assuming $\mathbf{PA}$ is consistent (3.25) is not true in N . For it were, if it would follow from Proposition 3.4.2 that $N \models (\exists y)\text{Prf}(\mathbf{p}(y), \ulcorner \perp \urcorner)$, which is true only if $\mathbf{PA}$ is inconsistent. $\square$

Proposition 3.4.3 shows that it not possible to extend the Arithmetic Soundness Theorem from $\mathbf{QLP}_0$ to $\mathbf{QLP}$. This is due to the fact that GJUG allows for the internalization of statements expressing consistency like (3.20) which are provable in $\mathbf{QLP}_0$. In fact, it is routine to check that the image of (3.23) under $(\cdot)^\circ$ is already false in N by employing the definition of $(\langle r(x)\forall x \rangle)^\circ$ given above (in particular, as long as $\mathbf{PA}$ is consistent, the number denoted by $g(\ulcorner r(x)^\circ \urcorner, \ulcorner x \urcorner)$ cannot denote a proof in P). It thus follows that GJUG is not an arithmetically sound rule of inference. Although this observation depends on the definition of $g(y_1, y_2)$, the general phenomenon persists even when we limit attention to theorems of $\mathbf{QLP}$ which do not contain the universal verifier symbol.

Proposition 3.4.4. Let $\mathcal{P}$ be axiomatically appropriate and let

$$\mathbf{QLP}^- = \{\varphi \mid \mathbf{QLP}(\mathcal{P}) \vdash \varphi \text{ and } \varphi \text{ does not contain the symbol } \langle \cdot \forall \cdot \rangle\}$$

Suppose also that $(\cdot)^\circ$ is a $\mathcal{P}$ -arithmetically valid interpretation and let $(\mathbf{QLP}^-)^\circ$ denote the image of all sentences of $\mathbf{QLP}^-$ under $(\cdot)^\circ$. Then $\mathbf{PA} \cup (\mathbf{QLP}^-)^\circ$ is inconsistent.

Proof. Note that by (3.24) and (3.25) $(\exists y)\text{Prf}(\mathbf{p}(y), \ulcorner (\exists y)\text{Prf}(\mathbf{p}(y), \ulcorner \perp \urcorner) \urcorner \rightarrow \perp) \in (\text{QLP}^-)^\circ$. It follows again by Proposition 3.4.2 that $\text{PA} \cup (\text{QLP}^-)^\circ \vdash (\exists y)\text{Prf}(\mathbf{p}(y), \ulcorner \perp \urcorner)$. But, as we have seen, $\text{QLP} \vdash (\exists x)x : \perp \rightarrow \perp$ and the arithmetic interpretation of this formula – i.e. $(\exists y)\text{Prf}(\mathbf{p}(y), \ulcorner \perp \urcorner) \rightarrow \perp$ – is thus in $(\text{QLP}^-)^\circ$. Putting these two facts together, we have $\text{PA} \cup (\text{QLP}^-)^\circ \vdash \perp$.

□

Turning finally to applications of the arithmetic interpretation of QLP_0 to the Knower Paradox, we may note first that it is an immediate consequence of the Arithmetic Soundness Theorem that QLP_0 is consistent with all statements whose images under all interpretations $(\cdot)^\circ$ are theorems for PA . Since the applying the Diagonal Lemma to the open sentence $\neg(\exists y_2)\text{Prf}(\mathbf{p}(y_2), y)$ is such a statement, we have the following:

Proposition 3.4.5. Let $\mathcal{P}$ be an axiomatically appropriate primitive term specification. Then

$$\text{QLP}_0(\mathcal{P}) \cup \{(\exists x_1)x_1 : (P \leftrightarrow \neg(\exists x_2)x_2 : P) \mid P \text{ a propositional letter}\}$$

is consistent.

Proof. Suppose for a contradiction that $\text{QLP}_0(\mathcal{P}) \vdash \neg(\exists x_1)x_1 : (P \leftrightarrow \neg(\exists x_2)x_2 : P)$. It follows by Theorem 3.4.1 that $N \models (\neg(\exists x_1)x_1 : (P \leftrightarrow \neg(\exists x_2)x_2 : P))^\circ$ for all $\mathcal{P}$ -arithmetically valid interpretations $(\cdot)^\circ$. From this it follows that

$$(3.26) \quad \neg(\exists y_1)\text{Prf}(\mathbf{p}(y_1), \ulcorner P^\circ \leftrightarrow \neg(\exists y_2)\text{Prf}(\mathbf{p}(y_2), \ulcorner P^\circ \urcorner) \urcorner)$$

is true in N . Note, however, that since $\text{PA} \supseteq \text{Q}$, $\text{PA} \vdash \delta \leftrightarrow \neg(\exists y_2)\text{Prf}(\mathbf{p}(y_2), \ulcorner \delta \urcorner)$

for some closed arithmetic sentence δ . Since PA also satisfies

$$\text{PA} \vdash \varphi \implies \text{PA} \vdash (\exists y)\text{Prf}(\mathbf{p}(y), \ulcorner \varphi \urcorner)$$

, it follows that $\text{PA} \vdash (\exists y_1)\text{Prf}(\mathbf{p}(y_1), \ulcorner \delta \leftrightarrow \neg(\exists y_2)\text{Prf}(\mathbf{p}(y_2), \ulcorner \delta \urcorner) \urcorner)$. But from this it follows that if $(\cdot)^\circ$ is any $\mathcal{P}$ -arithmetically valid interpretation such that $P^\circ = \delta$, then $N \models (\exists y_1)\text{Prf}(\mathbf{p}(y_1), \ulcorner P^\circ \leftrightarrow \neg(\exists y_2)\text{Prf}(\mathbf{p}(y_2), \ulcorner P^\circ \urcorner) \urcorner)$ in contradiction to (3.26). $\square$

A similar argument also suffices to demonstrate the result mentioned at the end of the prior section about the non-derivability of certain instances of U_q in QLP_0 .

Proposition 3.4.6. For all propositional letters P and axiomatically appropriate primitive term specifications $\mathcal{P}$, $\text{QLP}_0(\mathcal{P}) \not\vdash (\exists x_1)x_1 : ((\exists x_2)x_2 : P \rightarrow P)$

Proof. Suppose for a contradiction that $\text{QLP}_0(\mathcal{P}) \vdash (\exists x_1)x_1 : ((\exists x_2)x_2 : P \rightarrow P)$. By Theorem 3.4.1, it follows that $N \models ((\exists x_1)x_1 : ((\exists x_2)x_2 : P \rightarrow P))^\circ$ for all $\mathcal{P}$ -arithmetically valid interpretations $(\cdot)^\circ$. But as in the proof of Proposition 3.4.5, there exists a closed arithmetic sentence δ such that $\text{PA} \vdash$

$(\exists y_1)\text{Prf}(\mathbf{p}(y_1), \ulcorner \delta \leftrightarrow \neg(\exists y_2)\text{Prf}(\mathbf{p}(y_2), \ulcorner \delta \urcorner) \urcorner)$. Note that it also follows from the Arithmetic Soundness Theorem that since $\text{QLP}(\mathcal{P}) \vdash (\exists x)x : P \rightarrow P$ and thus the $N \models ((\exists x)x : P \rightarrow P)^\circ$ for all $(\cdot)^\circ$. If we again let $(\cdot)^\circ$ be such that $P^\circ = \delta$, we reach a contradiction by reasoning in the standard model in the same manner as Proposition 3.1.3. $\square$

Proposition 3.4.6 shows that without GJUG, QLP_0 is unable to derive the instance of $\mathbf{U}_q$ which we saw in the prior section is required in the proof of the Knower. Proposition 3.4.5 similarly shows that in the absence of this rule, we may consistently adjoin instances of the self-referential statements which rise give to the Knower. In a purely deductive sense, it thus seems that we can escape from the paradox by adopting QLP_0 as a basic framework for reasoning about knowledge and adjoining various self-referential schema involving justification as the context requires. The question remains whether this is a conceptually adequate resolution to the Knower Paradox relative to the sorts of desiderata which have been considered in the epistemic tradition. In the next section I will argue that not only can this resolution be motivated conceptually, but that a similar restriction on the internalization of reasoning universal generalization applied proofs is suggested by Gödel [37], [38] in a similar context.

3.5 GJUG, consistency and internalization

As I have mentioned previously, most commentators in the epistemic tradition attempt to resolve the Knower Paradox by arguing that either one of the epistemic axioms T or U or an internalization principle such as Nec ought to be rejected on the basis of the intended interpretation of $K(x)$. We have now seen that QLP derives T_q and U_q which are analogous to T and U . Similarly, the Lifiting Lemma of QLP serves a role analogous to Nec . Note, however, that none of these principles are basic to QLP but rather must be derived on the basis of putatively more fundamental principles about justifications and their combinatorial properties. In assessing the status of the Knower relative to QLP, we are thus presented with a larger set of options relative to which we might attempt to locate the source of the paradox.

The thesis which I will advance in this section is that the most objectionable principle on which the proof of the Knower relies is the Generalized Justified Universal Generalization rule [GJUG] mentioned above. We have now seen that without this rule we cannot derive the instance of U_q on which the Knower relies in QLP. As should now be evident, U_q is the image of the modal principal U under $(\cdot)^\exists$, whose image is the arithmetical principle U under $(\cdot)^*$. As such, the view I will advocate has the consequence that U is

indeed objectionable. But since the case I will advance is targeted at GJUG and not at U directly, it will provide a means of refining the arguments of other theorists who have located the source of the paradox either with U or variants of the Nec rule by which it can be derived from T.

As I mentioned in Section 1, the proposal to resolve the Knower by rejecting U represents a sort of consensus within the epistemic tradition. The *locus classicus* for this proposal is Anderson [1]. The proximal source of Anderson's rejection of this principle is an analogy between the knowledge predicate $K(x)$ and a formal provability predicate such as that suggested in (3.9). Anderson notes that on this interpretation there will be instances of U which will be false in the standard model: "If we rewrite $[\delta \leftrightarrow K(\neg\delta)]$... putting *Bew* everywhere for *K* ... the thing that's false for Q is the analogue of [U]." [1], p. 350. In order to motivate this remark, Anderson asserts that he is assuming that Q is capable of proving an appropriate version of the Second Incompleteness Theorem. In parallel to Proposition 3.4.3, it will then follow that $Bew(\ulcorner Bew(\ulcorner \perp \urcorner) \rightarrow \perp \urcorner)$ will be false in the standard model.

As we have seen above, however, this consideration constitutes at best an equivocal reason for giving up U in the context of the Knower. For not only is it unclear whether an appropriate version of the Second Incompleteness Theorem is indeed available for Q, but the motivation for adopting (3.9)

which analyzes $K(x)$ in terms of formal provability seems to be no stronger than that for adopting the QLP analysis (3.10) which does not commit us to thinking of justification as corresponding to formal provability in any particular system. It is thus interesting to note that Anderson also offers a positive argument in favor of a reformulated version of U. This argument is based on the familiar explanation of why the Gödel sentence of a theory like Q which is conventionally used in the proof of Gödel's First Incompleteness Theorem must not only be unprovable but also true.

This explanation runs as follows: 1) call the set of godel numbers of statements known by i K_0 , which we assume to contain Q, and to be consistent and recursively enumerable; 2) let T_0 be the deductive closure of K_0 and let γ be the Gödel sentence of T_0 – i.e. $T_0 \vdash \gamma \leftrightarrow \neg(\exists x)Proof_{T_0}(x, \ulcorner \gamma \urcorner)$ – and consider whether γ is provable in T_0 ; 3) if it is, then by the Σ_1 -completeness of T_0 , $T_0 \vdash (\exists x)Proof_{T_0}(x, \ulcorner \gamma \urcorner)$ and hence also $T_0 \vdash \neg\gamma$, contra to the assumption that this theory is consistent; 4) hence since γ isn't provable in T_0 , γ must be true since it expresses its own unprovability in T_0 . If we credit i with an ability to argue in this manner, it follows that he can learn that γ is true. But note that it follows from step 3) that $\gamma \notin K_0$ (as otherwise we would have $T_0 \vdash \gamma$). Thus if i could come to know γ by rehearsing the argument just described, he would thereby be enlarging the set of sentences K_0

which we took him to know initially to form some larger set K_1 containing γ .

Anderson takes this argument to motivate the proposal that knowledge is stratified into levels in a manner which may be characterized by adopting a sequence of knowledge predicates $K_0(x), K_1(x), \dots$ applicable to (the godel numbers) of sentences. Using such a sequence, we can formulate a stratified form of the principle U:

$$(U_i) \quad K_{i+1}(\ulcorner K_i(\ulcorner \varphi \urcorner) \rightarrow \varphi \urcorner)$$

Anderson motivates this principle on the basis of another familiar metamathematical argument: 1) let K_0 be the set of statements known by i which we again assume to contain $\mathbf{Q}$ and to be recursively enumerable and consistent; 2) suppose that the rules of $\mathbf{Q}$ preserve truth; 3) it hence follows that any sentence provable from K_0 is true. Since this reasoning can again presumably be carried out by i himself, it follows that i can come to know $K_0(\ulcorner \varphi \urcorner) \rightarrow \varphi$ (or even $K_0^*(\ulcorner \varphi \urcorner) \rightarrow \varphi$ where K_0^* is the deductive closure of K_0) for any φ . But in light of the Second Incompleteness Theorem, i 's knowledge of certain instances of this schema (e.g. with $\varphi \equiv \perp$) cannot correspond to sentences in K_0 . Thus if we take i to know these instances of T (as the argument suggests that we should), then they must fall under some larger set K_1 containing K_0 .

Anderson shows that there is a straightforward inductive argument which demonstrates that the theory T_ω formed by adjoining U_i for all $i \in \mathbb{N}$ to Q , T and a stratified version of the closure rule I is consistent. However, we need not concern ourselves with the details of this construction as it is possible to cite at least two conceptual difficulties with the hierarchical strategy in general. The first is the obvious analogy to Kripke's [92] familiar critique of Tarski's proposal to introduce a hierarchy of truth predicates to escape the Liar Paradox: since natural language (and presumably also our naive concept of truth) admits only a single, unstratified truth predicate, it seems unnatural to attempt to escape from the paradox by introducing a hierarchy of such predicates. The same observation appears to apply with equal or greater force to the everyday notion of knowledge. This is to say that neither in our everyday discourse about knowledge nor in familiar instances of reasoning about knowledge do we recourse to speaking of an agent knowing a proposition 'at one level' while simultaneously not knowing it another. To introduce such a distinction between levels of knowledge is thus to invite a variety of questions – e.g. why are the propositions considered by Anderson to be knowable at level $i + 1$ not knowable at level i by means other means than those he considers (e.g. by testimony)? how do we determine the level of a knowledge attribution made using our conventional unstratified

expression ‘knows that’? are there non-metamathematical examples which demonstrate that the hierarchy does not collapse? – to which our everyday intuitions appear to have no answer.

Pogliolesi [81] raises a related but potentially more serious objection to the hierarchical approach which is grounded in Anderson’s motivation for introducing the hierarchy itself. The essential observation is that both of his arguments for stratification equivocate between two different notions of knowledge. On the one hand, the extension of K_0 is fixed on the basis of the assumption that the extent of i ’s knowledge can be captured by a fixed, recursively enumerable theory, thus suggesting that knowledge is being analyzed in terms of formal proof. But in the course of these arguments, a case is made for the fact that i knows some statements outside of K_0 on the basis of an additional metamathematical argument. However, it is easy to see these arguments do not correspond to reasoning which can be conducted in K_0 . For instance, in the case of the first argument, in order to conclude that $\neg(\exists x)Proof_{T_0}(x, \ulcorner \gamma \urcorner)$ is true in the standard model requires us to reflect on the meaning of this statement in a way which cannot be captured by reasoning in $T_0 \supseteq K_0$. Similarly, the second argument requires that i can survey all possible derivations from K_0 and determine that they do not eventuate in a contradiction. But this again is not a task that can be

carried out in K_0 for well known reasons.

This not to say that the reasoning involved in these arguments is in any way defective or even not available to i in principle. Rather what the foregoing observations suggest is that the statements which Anderson argues fall under the extension of K_1 but not K_0 are provable in what Myhill [77] and Reinhardt [91] call the *absolute* sense – i.e. not in any particular formal system, but rather demonstrable via intuitively sound inferences from intuitively true axioms. If this is correct, then the concept of knowledge which K_1 is intended to analyze is different (perhaps quite radically) from that which K_0 is intended to analyze. As such, there seems to be no *a priori* reason why we should think they should obey similar logical principles.²²

One consequence of this is that while the considerations which Anderson adduces for the hierarchical strategy might provide some insight into the principles which should be contained in K_1 , they do little to fix the structure of the higher levels of the hierarchy. It is thus interesting to note that in assessing his proposal, other theorists have tended to agree with his isolation of U as problematic while failing to accept his argument that this should

²²In fact we might wish to forbid instances of K_i from occurring within the scope of K_j for $i > j > 1$ on the basis of the fact that it unclear what it would be means to prove (in the sense of K_i) some statement which itself involves the notion of absolute provability (in the sense of K_j). Note that Myhill [77] imposes a similar restriction on necessitation which blocks the derivation of statements with iterated applications of his absolute provability operator.

lead us to adopt a stratified version of this principle. For instance, both Égré [25] and Poggliolesi [81] suggest that the same considerations which Anderson cites can be used to argue for the appropriateness of adopting the propositional modal logic **GLS** as the appropriate medium for reasoning about the sort of knowledge at issue in the Knower. The simplest characterization of **GLS** is as the logic containing all of the theorems of **GL**, all instances of the axiom **T** and having modus ponens as its sole rule or inference. (In particular, **GLS** does *not* contain the rule **Nec**.) Since **GLS** contains all theorems of **GL**, it *ipso facto* has a means of producing self-referential statements via the De Jongh-Sambin Fixed Point Theorem. And since it also contains all instances of **T**, it is at least plausible to treat the notion formalized by the operator $\Box$ in this system as expressing a form of knowledge. In addition to this, a well known result of Solovay [99] demonstrates that the theorems of **GLS** correspond to those modal formulas φ such that φ^* is true in the standard model for all arithmetic interpretations $(\cdot)^*$. This in turn secures that **GLS** is consistent in the sense of not admitting any analogue of the Knower.

The component of the arithmetic theories relative to the paradox was originally framed which is missing from **GLS** is an analogue to the necessitation rule **Nec** or another principles like **I** which can be used to play a similar role. Note, however, **GLS** can also be characterized as the system containing

all axioms of **GL** and **T**, modus ponens and the rule **Nec** restricted to axioms of **GL** (cf. Boolos [7], p. 65-66). Thus **GLS** can be characterized so that it does contain a necessitation rule, although not one which is universally applicable in the sense of a normal modal logic. In particular, no instance $\Box(\Box\varphi \rightarrow \varphi)$ of **U** is derivable in this system unless φ (and therefore also $\Box\varphi$) is already derivable in **GL**. Égré and Poggliolesi thus claim that the source of the Knower lies with an unrestricted necessitation principle like **Nec**. Since **GLS** contains only a single modal operator, they are thus also able to endorse Anderson's argument against a one-level version of **U** without thereby also needing to consider the possibility of an iterated Knower involving knowledge predicates of higher type.

But beyond citing the argument of Anderson summarized above (which, as we have seen, depend crucially on the arithmetic features of **Q**), neither Égré nor Poggliolesi adduce any intrinsically epistemic considerations in support of restricting necessitation. Note, however, that the status of **U** with respect to **GLS** bears a strong resemblance to status of $\mathbf{U}_q$ with respect to $\mathbf{QLP}_0$. What I will now endeavor to show is that necessity of employing the **GJUG** rule to derive $\mathbf{U}_q$ from $\mathbf{T}_q$ in **QLP** can be taken to provide an argument against the acceptability of the former principal in the context of the Knower which can be motivated in terms of conceptual concerns about the nature

of justification rather than on the basis of the sorts of arithmetic analogies considered by Anderson.

I have already highlighted one concern which might lead one to be suspicious of $\mathbf{U}_q$. In particular, even if we accept that the concept of justification relevant to the Knower entails that every justification t which we can in some sense construct or inspect satisfies the explicit reflection principle $t : \varphi \rightarrow \varphi$, we might still resist the conclusion that this property holds for all justifications which fall under this notion. This is to say that we might think that we are justified in accepting each instance of this principle and might even believe that its universal closure – i.e. the schema $(\forall x)(x : \varphi \rightarrow \varphi)$ which is logically equivalent to $\mathbf{T}_q$ – to be true. But at the same time we might reasonably dissent from the claim that we *knew* the universally quantified statement to be true in the sense of possessing some sort of justification which verifies it irrefutably. But note that as we have seen in the derivation 5)-9) in Section 3, it is precisely the rule GJUG which allows us to infer from the fact that we possess justification for each instance of $x : \varphi \rightarrow \varphi$ that we thereby also possess justification for $\mathbf{T}_q$.

Making further sense of these observations inevitably requires that something additional be said about what notion of justification is relevant to the Knower. As I stressed in Chapter 1, the adoption of the language of QLP

is meant to be neutral between analyzing the notion of justification in three distinct ways: 1) the everyday sense of an evidential consideration which increases the degree of confirmation of a proposition for an agent; 2) the notion of informal proof in the sense of either non-formalized mathematical practice or the notion of absolute proof discussed by Myhill or Reinhardt; 3) formal proof in a particular axiomatic system. Since the first notion of justification is clearly incompatible with factivity, this notion is presumably not relevant to the setting of the Knower as all versions of the paradox we have seen rely on some form of the T axiom. But although I have argued that this context appears to be based on a notion of knowledge which requires proof of some sort, I have thus far avoided committing myself as to whether the relevant notion is that of formal or informal proof. For note on the one hand, to assume a formal analysis is likely to embroil us arithmetical considerations (e.g. the provability of Second Incompleteness Theorem in Q) which may reasonably be taken to be exogenous to the setting of the Knower. And on the other hand, Anderson's arguments might now lead one to suspect that in reality a mixture of intuitions about formal and informal contribute the principles assumed in Propositions 3.1.1, 3.1.2 and 3.1.3.

For this reason, let us finally suppose that the the notion of proof on which the Knower is based is that of informal mathematical proof as characterized

in Chapter 1. The question which we must now ask is whether the GJUG rule ought to be accepted as sound when the quantifiers of QLP are interpreted as ranging over such objects. Due to the imprecision in the notion of informal proof itself, this is not a question to which we can expect a completely formal answer. However, it is notable that a closely related question was considered by Gödel in the course of outlining the system of proto-explicit modal logic described at the beginning of Chapter 2.

One of goal of the paper [37] in which Gödel presented this system, was to consider whether a modal embedding similar to that considered in [35] could be employed to give a consistency proof for classical arithmetic and analysis. In order to do this, Gödel considers how it might be possible to provide a classical analysis of Heyting's [52] so-called proof interpretation of intuitionistic logic using a refinement of his [35] embedding of intuitionistic propositional calculus into **S4**. As described in Chapter 2, he proposed a language with proof quantifiers very similar to that of QLP. Adapting Gödel's notation to that of QLP, he asserts that within this system a statement of the form

$$(3.27) \quad a : (\forall x) \neg x : \perp$$

is provable. As Gödel observes, this can readily be seen to be an expression

of the provability of the consistency of the system he is considering, as it states that a is a proof that there is no proof of $\perp$. And in parallel to the conclusion he reached in [35] about the interpretation of the operator $\Box$ in **S4**, he cites the derivability of (3.27) as demonstrating that the quantifiers of the explicit modal system cannot be taken as ranging over proofs in a particular formal system.

This observation does not yet seem to be of any particular significance to the Knower Paradox. But note that taking $a = \langle r(x)\forall x \rangle$, we see that (3.27) is derivable in **QLP** via a familiar argument:

- | | | |
|----|---|-------|
| 1) | $\vdash x : \perp \rightarrow \perp$ | LP3 |
| 2) | $\vdash r(x) : (x : \perp \rightarrow \perp)$ | RQLP3 |
| 3) | $\vdash \langle r(x)\forall x \rangle : (\forall x)(x : \perp \rightarrow \perp)$ | GJUG |

We can now see that it follows by an easy variation on Proposition 3.4.6 that for no choice of a is the corresponding instance of (3.27) derivable in **QLP**₀. From this it follows that the use of the rule **GJUG** is essential to deriving this statement.

This observation is potentially of historical significance because although Gödel employs proof quantifiers, he does not state rules governing their introduction (in particular, he does not explicitly state rules analogous to either

UG or GJUG).²³ But in addition to this, it should also be noted that in the rest of [37] Gödel is concerned with stating adequacy conditions on what he refers to as ‘constructive systems’ as part of an attempt to provide a criterion for assessing which formal systems are suitable for providing consistency proofs which are of greater epistemic significance than that Gentzen’s consistency proof for arithmetic. Among these conditions is a requirement that the domain of objects over which the quantifiers on this system are intended to range are (as Gödel puts it) “surveyable (that is denumerable)” ([37], p. 91). Gödel’s concern appears to have been that if the objects over which we attempt to quantify in theory cannot be described in some uniform manner, then we lack justification for concluding that any property holds of them universally. He makes this point more explicitly in an earlier lecture [38] (from which parts of [37] are directly derived) in which he states that in a constructive setting, universal quantification is to be restricted to “infinite totalities for which we can give a finite procedure for generating all of their elements.” ([38], p. 51)

We may now ask whether the concept of constructive proof with which Gödel was concerned satisfies this requirement. The open-endedness of constructive proof (or more specifically our inability to identify it with provability in a given axiomatic system) is, of course, a commonplace of intuitionistic

mathematics.²⁴ And for this reason, Gödel explicitly rejected the possibility of a formalism which purported to quantify over all such proofs:

Heyting's axioms differ from [classical number theory] only by the fact that the substrate on which the constructions are carried out are proofs instead of numbers or other enumerable sets of mathematical objects. But by this very fact they do violate the principle which I stated before, that the word "any" can be applied only to those totalities for which we have a finite procedure for generating all their elements . . . Totalities whose elements cannot be generated by a well-defined procedure are in some sense vague and indefinite as to their borders. And this objection applies particularly to the totality of intuitionistic proofs because of the vagueness of the notion of constructivity. [38], p. 53

It is notable that both sorts of indeterminacy which Gödel claims to apply to constructive proof also apply *mutatus muntandis* to the notion of informal or absolute proof we are considering in relation in the Knower. For note that there is no reason to think that the sort of informal reflective methods described by Anderson represent anything like a closed or definable

²⁴E.g. "Of course, one is never sure that the formal system represents fully any domain of mathematical thought; at any moment the discovering of new methods of reasoning may force us to extend the formal system." Heyting [51], p. 5

class of techniques whereby i might come to extend his knowledge beyond the theorems of a particular axiomatic system. And thus although the notion of knowledge on which the Knower is based does seem to be correctly analyzed in terms of existential quantification over some notion of proof, there is thus no reason think that the domain of such proofs is precisely defined or even completely determinate.

With these observations in hand, let us finally return to the rule GJUG itself. Note that in the case under consideration the antecedent of this rule has the form $t(x) : \varphi(x)$ where x is a free proof variable and the conclusion has the form $\langle t\forall x \rangle : (\forall x)\varphi(x)$. The former expresses that $t(x)$ serves to justify $\varphi(x)$ (where this formula may be an open formula which states that a certain relation holds between the justification denoted by x and some potentially complex formula involving x). The conclusion, on the other hand, expresses that the justification denoted by $\langle t\forall x \rangle : (\forall x)\varphi(x)$ serves to justify the universal closure of $\varphi(x)$. In assessing the soundness of this rule in the context under consideration, we must seek to clarify two points: 1) what does it mean for a statement of the form $t(x) : \varphi(x)$ with x free to be derivable in QLP? 2) is there always a uniform way of going from a proof of $\varphi(x)$ to a proof of its universal closure?

A typical case of the first sort is given by the derivation of the formula

$s(x) : (\exists y)[x : \varphi \rightarrow y : x : \varphi]$ in Chapter 2.3.2. Here, $s(x)$ denotes a complex proof term, built up from primitive proof terms and proof operators which encodes a derivation of $(\exists y)[x : \varphi \rightarrow y : x : \varphi]$. Note that despite the fact that x occurs free in this formula, the structure of $s(x)$ will reflect the structure of a proof of $(\exists y)[x : \varphi \rightarrow y : x : \varphi]$ in QLP which is uniform in x in the sense that were we to substitute any proof term u for x , $s(u)$ would then denote a proof of the corresponding closed formula $(\exists y)[u : \varphi \rightarrow y : u : \varphi]$. According to the motivation of the GJUG rule offered in Chapter 2, this fact can be taken to justify our acceptance of $(\forall x)(\exists y)[x : \varphi \rightarrow y : x : \varphi]$ and the universal verifier term $\langle s(x)\forall x \rangle$ is accordingly introduced as a means of naming the proof corresponding to this justification.

Although the derivation 1)-3) given above has the same overall form, the application of GJUG at step 3) is harder to motivate as being clearly valid. Although line 1) is simply an instance of the QLP reflection axiom, one might already be troubled by how to interpret the free variable x . In particular, although we might be tempted to interpret the variable x as being universally bound, this may not make sense in the context where we are calling into question whether there is a determinate domain of justifications over which QLP quantifiers range. Note, however, that we do presumably accept that for every informal proof t which we can construct, if t is a proof of φ , then φ

must be true. And thus inasmuch as we take this to be an intrinsic feature of the notion of informal proof itself (as, again, the setting of the Knower appears to require), then there seems to be nothing wrong with expressing this fact via the use of a free variable. Line 2) results from applying the rule of axiom necessitation to line 1), which in the current case simply means introducing a parameterized family of names $r(t_1), r(t_2), \dots$ to serve as names for justifications of $t_1 : \perp \rightarrow \perp, t_2 : \perp \rightarrow \perp, \dots$

Turning now to step 3), we can see that there is a potential problem standing in the way of securing the applicability of GJUG. For in this case, rather than encoding a template for constructing a proof which itself serves to justify $t : \perp \rightarrow \perp$, $r(t)$ simply functions as a name for a primitive justification for this formula. Thus even if we are prepared to accept $t : \perp \rightarrow \perp$ for every informal proof t we have thus far encountered or can envision constructing, the structure of the proof term $r(x)$ does not itself allow us to project whatever credence we have in the factivity of informal proof to new cases. Given the open-endedness which the concept of informal proof appears to possess, it thus seems at best optimistic to claim that we possess a justification for the claim expressed by the universal statement $(\forall x)(x : \perp \rightarrow \perp)$. For if t denoted such a justification, our acceptance of this claim would *ipso facto* also commit us to the fact that t itself is factive. Thus rather than

merely recording the fact that we have some non-demonstrative reason to believe that all informal proofs are factive, the existence of such a t would entail that $(\forall x)(x : \perp \rightarrow \perp)$ could not fail to be true. But this seems like a far stronger claim than we could ever possibly have evidence for (let alone irrefutable evidence), even if it turned out to be true.

The obvious upshot of these considerations is that we appear to have a reason to reject the legitimacy of the application of GJUG at line 3) of the foregoing derivation which is based in what can reasonably be described as conceptual considerations about the background notion of informal proof. Note, however, that although 3) serves as a consistency statement, nothing about the foregoing argument has depended on this. Thus the same argument appears to call into doubt the legitimacy of using GJUG to derive any instance of $\langle r(x)\forall x \rangle : (\exists x)x : \varphi \rightarrow \varphi$ for any φ which is not already provable in QLP. As such, this argument can also be taken to cast into doubt the epistemic status of $\mathbf{U}_q$. If this is the case, then it seems that we have indeed found an epistemic rationale for blocking the derivation of the Knower in QLP, even if we assume the existence of provably self-referential statements in the manner indicated by Proposition 4.5. Such a resolution appears to share all of the virtues of the proposal of Poggiolesi and Égré while also having the added benefit of providing an explanation of why unrestricted

necessitation fails when we attempt to analyze the Knower in a traditional modal language.

Chapter 4

QLP and the Knowability Paradox

4.1 Introduction

The Knowability Paradox purports to show that from the controversial but not patently absurd hypothesis that all truths are knowable, it is possible to derive the implausible conclusion that all truths are known. These statements can be expressed schematically as follows:

(KP) For all φ , if φ is true, then it is possible to know φ .

(Om) For all φ , if φ is true, then φ is known.

(KP) is traditionally referred to as the *Knowability Principle*. It is conventionally formalized in a propositional modal language $\mathcal{L}_{LK}$ containing operators $\Diamond$ (*it is possible that*) and K (*it is known that*). In this language, (KP) takes the form

(KP1) For all φ , $\varphi \rightarrow \Diamond K\varphi$.

(Om) is accordingly formalized in $\mathcal{L}_{LK}$ as

(Om1) For all φ , $\varphi \rightarrow K\varphi$.

On the interpretation of the modal operators just offered, (Om1) is presumably false. For surely as a matter of empirical fact, we are non-omniscient in the sense that

(NonOm) For some ψ , ψ is true but not known.

This claim is naturally rendered in $\mathcal{L}_{LK}$ as

(NonOm1) For some ψ , $\psi \wedge \neg K\psi$.

Suppose we now assume that the knowledge operator K is factive (i.e. $K\varphi \rightarrow \varphi$) and distributes over conjunction (i.e. $K(\varphi \wedge \psi) \rightarrow (K\varphi \wedge K\psi)$) and that provability implies necessity (i.e. $\vdash \varphi \therefore \vdash \Box\varphi$). If we now take (NonOm1) as a premise for reductio for some fixed ψ , then the corresponding instance of (Om1) can be derived via the following argument:

- | | |
|--|---|
| 1) $\psi \wedge \neg K\psi$ | NonOm1 |
| 2) $(\psi \wedge \neg K\psi) \rightarrow \Diamond K(\psi \wedge \neg K\psi)$ | KP1 (taking $\varphi \equiv \psi \wedge \neg K\psi$) |
| 3) $\Diamond K(\psi \wedge \neg K\psi)$ | 1), 2) |

4)	$K(\psi \wedge \neg K\psi)$	premise for reductio
5)	$K\psi \wedge K\neg K\psi$	distributivity of K
7)	$K\psi$	5)
8)	$K\neg K\psi$	5)
9)	$\neg K\psi$	factivity for K
10)	$\neg(K(\psi \wedge \neg K\psi))$	4), 5)-8)
11)	$\Box\neg(K(\psi \wedge \neg K\psi)) =_{df} \neg\Diamond(K(\psi \wedge \neg K\psi))$	necessitation 10)
12)	$\neg(\psi \wedge \neg K\psi)$	1), 3) - 11)
13)	$\psi \rightarrow \neg\neg K\psi$	12) (intuitionistic logic)
14)	$\psi \rightarrow K\psi$	13) (classical logic)

The observation that all instances of (Om1) can be derived from the assumption of (KP1) in the manner just described has come to be known as the the Knowability Paradox. It is conventionally taken to be paradoxical in virtue of the fact that while (KP) is not clearly false, it entails an apparent falsehood on the basis of what appear to be weak background assumptions about the behavior of the knowledge and necessity operators. As such, most modern commentators view the Paradox as casting into doubt any view which embraces any principle connecting truth and knowability resembling (KP).

This is not, however, the context in which the significance of statements of the form (NonOm1) was observed. In [27], Frederic Fitch attempted to formulate an analysis of the notion of value which (for reasons which need not concern us here) would have been trivialized by the existence of unknowable truths. Fitch noted that the assumption of (NonOm1) for a particular proposition ψ implies the existence of a truth χ which cannot be known. This follows by taking $\chi \equiv \psi \wedge \neg K\psi$ and carrying through steps 4) - 11) in the preceding derivation. These steps thus function not as a reductio of (KP), but rather as a demonstration that in the presence of the other assumptions about the K and $\Box$ operators mentioned above, the existence of an unknown truth entails the existence of a necessarily unknown (or *unknowable*) truth.¹

I will argue below that derivation 4) - 11) anticipates what we will present as a positive result about the existence of non-constructed proofs. But this is not the manner in which the complete proof 1) - 14) – which we will subsequently refer to as the *Fitch derivation* – is viewed in the contemporary literature on the Knowability Paradox. This can be taken to begin with Hart and McGinn's [49] rediscovery of the Fitch derivation and their subsequent

¹Fitch attributes this observation to an anonymous referee, who subsequent investigation has revealed to be Alonzo Church. Documentary evidence recently discovered by Salerno [93] also suggests that the proof Church had in mind was essentially the one given here as 4)-11). Note, however, that what this derivation appears to show on its own is that we cannot know that there is a unknown truth, not that one must exist categorically.

recasting of it as an antinomy for a view they refer to as *verificationism*. As understood by Hart and McGinn, “verificationism” describes a class of views according to which the notion of truth is subjected to a so-called *epistemic constraint* – i.e. a principle which states that in order for a statement to be true, it must satisfy some sort of criterion which makes substantial use of epistemic notions such as verification, knowledge or provability. I will follow Dummett’s (e.g. [19]) terminology in referring to an opponent of verificationism as a *realist* – i.e. a theorist who holds that the truth value of a statement is fixed independently of our capacity for knowing or verifying it.

Hart and McGinn suggest that (KP) can be taken to be a schematic formulation of verificationism. Although we will take issue with this claim below, a statement closely resembling this principle can indeed be found in Dummett’s [22] “What is a theory of meaning? (II).” Hart and McGinn thus argue that the Fitch derivation can be taken to show that verificationists must also accept (Om). But on the interpretation offered above, this principle will presumably be rejected by both realists and verificationists alike. On this basis, Hart and McGinn propose that the Fitch derivation ought to be taken as a *reductio* of verificationism.

4.1.1 A proof-based approach to the Paradox

My interest in this chapter is not in defending verificationism directly, but rather in demonstrating that the Knowability Paradox poses no threat to the views of practicing verificationists inasmuch as they are proponents of intuitionistic logic and mathematics. The potential connection between the Paradox and intuitionism derives from the fact that intuitionists have traditionally sought to define truth in terms of epistemic notions in a manner which is at least reminiscent of (KP). Standard introductions to intuitionism (e.g. [105], [17]) thus often begin by laying down that the assertion of a logically unstructured mathematical statement P should be understood to be equivalent to the claim that we are able to prove P . We may thus tentatively formulate the fundamental intuitionistic conception of truth for such statements as

(4.1) P is true if and only if P is verifiable.

subject to the understanding that to say P is verifiable is to say that we can construct a certain kind of proof of P .

The concern raised by the Knowability Paradox arises not from the fear that it undermines a specific formulation of verificationism such as Dummett's anti-realism, but rather that it might infect the logical substructure

of an entire species of views which adopt some version of (4.1). To appreciate the relevant concern, it is important to realize not only that (4.1) approximates the sort of epistemic limitation which intuitionists wish to impose on truth, but also that such an analysis has traditionally been taken to provide the basis for an intuitionistic understanding of the logic connectives as discussed briefly at the beginning of Chapter 2. In particular, it is generally accepted that a proper intuitionistic interpretation of the connectives must conform to (4.1) in the sense that it should specify the conditions under which a logically complex statement is verifiable in terms of the conditions under which its constituents are verifiable.

Within the context of the now standard Brouwer-Heyting-Kolmogorov [BHK] interpretation of the intuitionistic connectives, the notion of verification is analyzed in terms of the notion of *constructive proof*. The exact nature of such proofs has, of course, long been a matter of controversy within intuitionism (cf., e.g., [111]). However, the BHK interpretation itself merely seeks to specify what would constitute a proof of a logically complex statement in terms of the so-called *proof conditions* of its constituents. This gives rise to the familiar inductive characterization of such conditions whose clauses are typified by

(BHK_∧) a proof of $\varphi \wedge \psi$ is constituted by a proof of φ and a proof of ψ

(BHK $\rightarrow$) a proof of $\varphi \rightarrow \psi$ is constituted by a method which allows us to transform any proof of φ into a proof of ψ

By giving clauses of this sort for each of the propositional connectives and quantifiers, the BHK interpretation may be taken as specifying the meaning of a logically complex statement in terms of the proof conditions of its constituents. It thus follows that if intuitionistic truth is to be understood as verifiability, and verifiability is to be understood in terms of constructive provability, then for an arbitrary statement φ , intuitionists will accept the principle

(KP2) φ is true if and only if φ is constructively provable.

where the sense of the right-hand side is fixed according to the structure of φ by the BHK interpretation. Note, however, that to say that φ is provable, is presumably to say nothing more than that there exists a constructive proof of φ . As such, I argue below that the sort of epistemic constraint on truth accepted by contemporary verificationists is most faithfully expressed by a principle of the form

(KP3) For all φ , if φ is true, then there exists a proof of φ .

It might first appear that (KP1) and (KP3) correspond to very different means of expressing an epistemic constraint on truth. And as such, one

might reasonably wonder what relationship the latter schema bears to the Knowability Paradox in both a rhetorical and a logical sense. With respect to the former issue, one of my goals in this chapter will be to argue that (KP3) does indeed more accurately express the views of theorists who are the presumptive targets of the Knowability Paradox than does (KP1). But if this is so, then one might worry that it is possible to formulate an analogous paradox premised on (KP3). In order to see how this might be possible, we must inquire as to whether the notion of *knowledge* which figures in (NonOm) should be understood relative to the proof-based analysis of knowability just outlined. As I will argue in more detail below, for a verificationist, the failure of omniscience is most naturally understood in terms of the existence of true but *unproven* statements. The existence of such statements can be expressed as follows:

(NonOm2) For some ψ , ψ is true and no proof of ψ has been constructed.

Note that (NonOm2) will presumably be accepted by realists on the basis not only of formally undecidable statements, but also on the basis of the presumption that there are simply many mathematical statements which, while true, we have simply not bothered to derive. However, gauging the likely attitude of intuitionists towards this statement is less straightforward. One

might initially think that acknowledging the existence of true but unproven statements appears to stand in direct conflict with Brouwer’s dictum that “there are no unexperienced truths”[12]. But it is also acknowledged that intuitionistic truth is a subtle notion and that there may not even be complete agreement among important figures about its proper analysis.² Taking this into account, I will argue below that there is a straightforward argument based on assumptions that would be accepted by most intuitionists – and in particular the theorists with whom I will be concerned – to the effect that (NonOm2) ought to be accepted.

The question thus naturally arises whether, when taken in conjunction with (KP3), (NonOm2) entails any schema parallel to (Om) which is in conflict with common intuitions about our everyday concepts of truth and knowledge. In the framework I will adopt, such a schema would take the form

(Om2) For all φ , if φ is true, then we have constructed a proof of φ .

It will be the second substantial goal of this chapter to show that no statement of this kind is derivable from (KP3) and (NonOm2) via any means by which either a realist or a verificationist would consider legitimate. In particular, I will argue that there is no way to reconstruct the reasoning of Fitch’s original

²Cf., e.g., [89] for discussion of different views of truth within intuitionism.

derivation to derive (Om2) from (KP3) and (NonOm2). I will additionally provide a consistency proof which demonstrates that there is no other means of deriving a contradictory statement from these principles.³

I will endeavor to do this in a way which is as precise and rigorous as possible. Since the reasoning of the Fitch derivation is reasonably involved, we will have to find a means of formalizing (KP3) and (NonOm2) in a language in which such principles can be stated directly. This task poses two distinct challenges. On the one hand, there is no way of formulating these principles in a conventional modal language such as $\mathcal{L}_{LK}$ as this language contains no means of quantifying over proofs. And on the other, once the relationship between intuitionism and verificationism has been highlighted, one is immediately drawn to ask whether the use of classical logic at the last step of the original derivation of the Paradox is legitimate. I will now briefly discuss both of these matters before turning to an in-depth analysis of the views of the verificationists and the technical exposition of the system I will

³The closest anticipators in proposing such a resolution to the Paradox are Cozzo [15] and Hand [46], [47]. I will consider Hand's proposal explicitly in Section 3.1. However, since there will no convenient point to return to Cozzo's proposal, I will simply note that on the basis of concerns similar to those I will discuss in Section 2, Cozzo argues that (KP) should be formalized using a quantifier over what he refers to as *ideal verifications*. The (KP)-like principle which he formulates thus mirrors (KP3) in form. However, Cozzo also elects to treat knowledge as an unanalyzed propositional operator as opposed to analyzing it in terms of our failure to have constructed a proof of ψ in the manner of (NomOm2). For this reason, it is difficult to directly compare his approach to the Paradox to the one adopted here.

use to represent them.

4.1.2 Knowledge, proof and explicit modal logic

In order to fully understand the significance which the Knowability Paradox holds for verificationism, it is necessary to analyze the particular notions of knowledge and knowability relative to which practicing verificationists will assent to (KP) and (NonOm). I have already taken a step in this direction by proposing that (KP3) and (NonOm2) better approximate the sense which such theorists will assign to these principles than do (KP1) and (NonOm1). Implicit in these proposals are the claims that verificationists will understand the notions of knowledge and knowability as follows:

- (4.2) a) φ is known if and only if a proof of φ has been constructed
 b) φ is knowable if and only if there exists a proof of φ

(4.2a,b) may seem unexpected as analyses of the familiar epistemic concepts which the Knowability Paradox is conventionally framed as drawing into conflict. However, subject to their assent to (KP) itself, it is not hard to see why these analyses represent the perspective on knowledge and knowability which the verificationists are likely to hold.

For note that there is no reason to think that verificationists will not agree with many realists in holding that the everyday notion of knowledge

can be analyzed in the traditional manner as a species of justified, true belief. Like a realist, a verificationist may acknowledge that such an analysis is incomplete in the sense that it fails to explain why certain intuitively admissible forms of justification are sufficient to produce knowledge while others are not. However, unlike many realists, verificationists will presumably be in possession of a reasonably specific analysis of justification in the form of their preferred notion of constructive proof. Such proofs are typically taken to satisfy the following properties:

- I) A constructive proof is a structured entity $\mathcal{D}$ constituted by a sequence (or tree) of statements or propositions, one of which is identified as its conclusion $Concl(\mathcal{D})$.
- II) Such proofs may be constructed by epistemic agents and if an agent i constructs a proof $\mathcal{D}$ such that $Concl(\mathcal{D}) = \varphi$, he is thereby justified in believing φ .
- III) If there exists a constructive proof $\mathcal{D}$ such that $Concl(\mathcal{D}) = \varphi$, then φ is true.

If we take these properties of constructive proofs into account, then it is not hard to see why verificationists will be likely to accept the analyses given in (4.2a,b). For recall that the Knowability Paradox turns on the fact that to

say that φ is *knowable* is equivalent to saying that φ is possibly known. But now note that according to the justified-true-belief analysis of knowledge, to say that φ is possibly known is equivalent to saying that there possibly exists some sort of justification which would serve to verify φ . But according to the verificationists, the notion of justification should be understood in terms of that of constructive proof. And from this it follows that to say φ is knowable for a verificationist is equivalent to saying that there exists a constructive proof with conclusion φ , wherein the existential quantifier is taken to range over not only proofs which are actual in the sense of having been constructed, but also over those which are merely potential in the sense that they could be constructed.

As I discussed above, there presumably are statements ψ which are knowable in the sense just described but which we do not know to be true. Note, however, that if we have constructed a proof of ψ , by II) we would be justified in believing it. And since by III), we would then also be assured that ψ was true, the justified-true-belief analysis would then imply that ψ is known. As such, the fact that ψ is not known in the envisioned case must presumably be explained in terms of the fact that no proof of ψ is available to us. Since proofs are the sort of things we construct, this in turn must be explained by the fact that we have not constructed a proof of ψ . And from this, it appears

that for a verificationist to accept that ψ is known, it is sufficient to require both that there is a proof of ψ and also that this proof has been constructed.

The foregoing considerations do not, of course, demonstrate that (4.2a,b) are unassailable as analyses of our everyday concepts of knowledge and knowability. For note that III) corresponds to the condition that construct proofs are a factive form of justification. If it is required that in order for an agent i to know φ , i must possess such a proof, it follows that we will be unable to account for cases in which we might credit i with knowing φ on the basis of a form of evidence which does not possess this property. This will paradigmatically be the case, for instance, when φ is an empirical proposition, and the justification which i possesses for φ is either premised on other empirical hypotheses or perceptual judgements.⁴

This concern aside, if we accept that (4.2a,b) serve at least as general templates for the form which the verificationists' analyses of knowledge and

⁴(4.2a) thus highlights how wide the rift is between the accounts of knowledge and justification which are likely to be favored by verificationists and those employed in mainstream epistemology. For contemporary verificationists do not think of (KP) as an isolated principle, but rather view it as a consequence of their belief that a theory of meaning should be modeled on the sort of proof theoretic explanations of the intuitionistic connectives discussed above. But the notion of proof which is most directly applicable to such an interpretation is that of a formal derivation in intuitionistic logic or mathematics. If the verificationists wish to formulate a theory of meaning which is applicable to empirical statements, they will presumably have to broaden their understanding of constructive proof to include demonstrations which can verify empirical statements. The question of how an appropriate notion of empirical demonstration might be extrapolated from the traditional notion of constructive proof has been discussed by Dummett (cf., e.g., [22], [21]) and Prawitz [85].

knowability must take, it follows that any formal language in which we might hope to represent their views must itself allow us to speak directly of proofs and their relationship to sentences. It is thus natural to employ the language of QLP, in which case (KP3) is naturally formalized as

(KP4) For all φ , $\varphi \rightarrow (\exists x)x : \varphi$.

However, the adoption of a language with proof quantifiers alone is not sufficient to distinguish between statements which are provable and those whose proofs have been constructed. In order to express this distinction, we will add to $\mathcal{L}_{\text{QLP}}$ a predicate $E(x)$ of proofs with the intended interpretation

(4.3) $E(t) \iff$ the proof denoted by t has been constructed⁵

Statements of the form $E(t)$ will serve as well-formed formulas in a system called QLPE. In the language of this system, (NonOm2) can naturally be formalized as

(NonOm3) For some ψ , $\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]$.

I will defer the formal exposition of QLPE until Section 3 where we will pursue the two technical goals announced above – i.e. 1) to show that the QLPE formalization of (Om2) (i.e. (Om3) For all φ , $\varphi \rightarrow (\exists x)[E(x) \wedge x : \varphi]$)

⁵A similar device was employed by Parikh [78] for expressing that a natural number has been constructed.

cannot be derived from (KP4) and (NonOm3); 2) to show that the latter two principles are formally consistent. Taken together, I will argue that these results demonstrate that the Knowability Paradox poses no threat to at least one plausible means of formulating verificationism. This caveat is necessary for several reasons. For although I will argue that these formalizations represent reasonable reconstructions of the core tenets of contemporary verificationism, the proponents of this view are not in complete agreement about all the issues which pertain to the proper formalization of these principles. These differences do not arise at the level of their endorsement of a principle like (KP3) which attempts to analyze truth in terms of proof existence. Rather, as we will see in Section 2, they come out with respect to the significance which the verificationists assign to the notion of constructive proof itself.

As we will see below, (KP4) and (NonOm3) allow us to derive the conclusion that there exist non-constructed proofs. The status of such “potential proofs” – i.e. derivations which have not been carried out by any actual epistemic agents but which are potentially constructible in the sense of falling under some general characterization of constructive proof – has been debated in the literature of contemporary verificationism. Although the notion of potential proof has been embraced by Prawitz and Martin-Löf, the status of

such entities is contentious. In particular, Dummett has repeatedly suggested that the acknowledgement that there exist merely potential proofs represents a substantial concession to realism. In Section 4 I will endeavor to show that acknowledging the existence of such proofs need not be taken to have the consequences which are of concern to Dummett. As such, it seems that the real significance of the Knowability Paradox is not to undermine verificationism, but rather to illuminate a way in which this view is committed to the existence of potential proofs.

Before embarking on these tasks in earnest, it will finally be useful to highlight one other aspect of the approach adopted here which differs from that of some other commentators on the Paradox. As was mentioned in passing above, the original Fitch derivation employs classical logic in the transition from step 13) to 14). From the time of its rediscovery by Hart and McGinn, this derivation has repeatedly been used as a means of critiquing verificationism. But since verificationists generally reject classical logic, it is natural ask to whether the very theorists against whom these critiques are targeted must accept the conclusion of the Paradox. This issue has influenced the literature on the Paradox in a variety of ways. For instance, in his original article, Hart [48] acknowledges that a case can be made that intuitionists should simply accept the principle (Om) which he otherwise

argues should be taken as a reductio of verificationism.⁶ Hart remarks that he finds this resolution untenable since, understood informally, (Om) expresses a principle which is radically out of keeping with our ordinary intuitions about knowledge. And as such, he proposes that the Paradox should be taken as a genuine antinomy for verificationism.

Williamson [113], on the other hand, attempts to intervene on behalf of the verificationists by arguing that they should reject the classically mediated step. This would save them from having to accept (Om) at the expense of accepting step 13) (i.e. $\varphi \rightarrow \neg\neg K\psi$). However a variety of theorists have objected that 13) is already an embarrassing consequence of verificationism. And in addition, Percival [79] has shown how to formulate a similar paradox based on the existence of undecided statements which uses only intuitionistically valid reasoning.

My ultimate view will be that the debate over the use of classical logic in the Fitch derivation represents something of a red herring. For in order for the Paradox to be deployed by realists as an argument against verifica-

⁶For note that if we interpret the conditional in (KP1) according to (BHK $\rightarrow$), then (Om1) appears to report the innocuous finding that were to possess a proof of φ then we could uniformly construct a proof of $K\varphi$ (presumably by reflecting on our original proof). This observation suggests that intuitionists possess a natural response to the Paradox – i.e. they should simply accept (Om). I will henceforth refer to this as the *bullet biting* resolution of the Paradox. Hart’s negative appraisal notwithstanding, some commentators have proposed that intuitionists ought to adopt such a resolution – e.g. [71].

tionism, it must presumably be formulated in a language which the realists can themselves understand. However, it is open to a realist to claim that he does not understand the meanings of the intuitionistic connectives. This consideration appears to favor the adoption of a classical background logic for use in assessing the consequences of (KP) and (NonOm). Verificationists can, of course, reply that they are analogously unable to understand the classical meanings of negation and implication which underwrite the legitimacy of the inference from 13) to 14). And for this reason, it should be left open for them to adopt the bullet biting argument described in note 6 for accepting (Om). But in taking this step, verificationists do indeed put themselves in the position of embracing a principle which is in strong conflict with our everyday intuitions. In particular, such a step may seem overly drastic, as it appears to undermine a verificationist's ability to express his views to the realist using a common language.

There is a long tradition of providing classical interpretations of intuitionistic logic and mathematics in a language which classical theorists can understand. It is often proposed that such interpretations can be used as a means of overcoming communication breakdowns of the sort just described. One of the earliest such attempts was Gödel's use of the operator $\mathcal{B}$ to interpret IPC in **S4** as described above. (KP3) itself can also be understood

as part of an attempt to interpret the notion of intuitionistic truth in terms which a realist can presumably understand – i.e. those of proof and existence. This analysis may be partial since the significance which an intuitionist attaches to the occurrence of “there exists” in (KP3) may itself differ from that which it is assigned by a realist. However, it seems that verificationists need not balk at the application of classical logic to statements in the language of QLPE for purposes of reasoning about the Paradox. For such a language need not be understood as the native tongue of a verificationist, but rather that of an idiom which he is temporarily adopting in order to thwart a threat which has been imposed from outside.

4.2 Neo-verificationism, knowledge and proof

My primary goal in this section is to argue that the interpretations of the premises of the Knowability Paradox – i.e. (KP4) and (NonOm3) – proposed above have a strong basis in the writings of contemporary theorists who describe themselves (or are often described) as verificationists. As mentioned above, I will take Michael Dummett, Dag Prawitz and Per Martin-Löf to be the paradigmatic exemplars of this class. These theorists have all famously endorsed versions of an epistemic constraint on truth in the sense explained above. However, the application of the term “verificationism” to describe

their views is at least slightly problematic due to the prior application of this term to describe the views of Ayer and the theorists of the Vienna Circle.⁷ To avoid any confusion on this issue, I will henceforth adopt the term “neo-verificationism” not only to differentiate the views of Dummett, Prawitz and Martin-Löf from these earlier theorists, but also to demarcate several positive factors which serve to distinguish their views.⁸

Perhaps the most significant feature which the neo-verificationists have in common that is they seek to identify the meaning of a statement not with its classical truth conditions but rather with some notion of a verification condition. In this setting, the notion of a verification is taken to generalize that of a constructive proof which is also “direct” in (roughly) the sense that its combinatorial structure reflects the logical structure of its conclusion. The notion of a direct verification is paradigmatically formalized via the notion of a canonical or normal proof in a Gentzen-style natural deduction system. In this context, the introduction rules of such a system are taken as definitionally giving the meaning of the logical constants, while the elimination rules are justified on the basis of their structural features relative to the introduction rules. For present purposes, it will be this shared intellectual legacy –

⁷Martin-Löf [70], p. 34 has been most explicit about drawing attention to the historical basis (and ultimate clash) of these applications of the term “verificationism.”

⁸This term has also been employed by Usberti [107] in a similar manner.

which notably differs from that of the early twentieth century verificationists both with respect to its connection with intuitionism and also with subsequent work in proof theory in the tradition of Gentzen – which unify the views of the neo-verificationists.

Although it is the neo-verificationists who are the presumptive targets of the Knowability Paradox, the literature on this subject interfaces only indirectly with the aspects of their views I have just described. In this latter-day setting, Dummett’s paper “What is a theory of meaning? (II)” [22] can be taken as the canonical exposition of the neo-verificationist position.⁹ This paper contains the following formulation of an epistemic constraint on truth:

(K) If a statement is true, then it must be in principle possible to know
that it is true. ([22], p. 61)

Principle (K) clearly proposes to analyze truth in terms of the notions of knowledge and possibility. And thus this statement appears to provide a clear basis for schematizing neo-verificationism by using propositional knowledge and possibility operators of the sort employed in (KP1).

In order to understand the role which (K) plays with respect to Dum-

⁹As we have seen, Dummett appears to be the target of the paper [48] in which Hart “rediscovers” (and repurposes) Fitch’s derivation. For although [22] is not explicitly cited as the source for the view which Hart seeks to critique, it contains a discussion of verificationism which is similar to the view described in this paper.

mett's views as a whole, it should be kept in mind that he repeatedly cites the intuitionistic understanding of the logical connectives as a paradigm for the form which a theory of meaning should take. From this it follows that the notion of proof must be assigned a central role within such a theory. For the meanings of the intuitionistic connectives are standardly explained via the BHK interpretation which seeks to specify the meaning of a statement in terms of its proof conditions.¹⁰ On this basis, Dummett proposes within [22] itself that to grasp the meaning of an arbitrary statement is to grasp what would count as a proof of it:

The intuitionistic explanations of the logical constants provide a prototype for a theory of meaning in which truth and falsity are not the central notions. The fundamental idea is that a grasp of the meaning of a mathematical statement consists, not in a knowledge of what has to be the case . . . for the statement to be true, but in an ability to recognize, for any mathematical construction, whether or not it constitutes a proof of the statement; an assertion of such a statement is to be construed, not as a claim that it is true, but as a claim that a proof of it exists or can be constructed. ([22], p. 70)

¹⁰It may also be noted here that Dummett, Prawitz and Martin-Löf all think that the

This passage is a typical illustration of the way in which the neo-verificationists propose to ground a general theory of meaning in an analysis of the relationship between a statement and the proofs which serve to verify it. General formulations of this view inevitably involve quantification over proofs in the sense that in order to grasp the meaning of φ , it is asserted that we are able to ascertain for *any* proof whether it constitutes a proof of φ . The prior passage appears to suggest that such quantification can be used to eliminate the need for the notion of truth within a theory of meaning. However at other times Dummett argues instead that truth should be *defined* in terms of proof – e.g.

A sentence may then be said to be true if and only if there exists a construction that constitutes a proof of it. ([19], p. 234)

[A]n understanding of a mathematical statement consists in the capacity to recognize a proof of it when presented with one; and the truth of such a statement can only consist in the existence of such a proof. ([17], p. 4)

analysis of meaning based on canonical proof described above should be thought of as a further refinement of ideas already implicit in the BHK interpretation – cf. [18], [82]. However, since the specific features which they assign to canonical proofs will not play a substantial role here, I will generally not bother to issue caveats which flag the (often very complex) ways in which this notion contributes to their views.

In drawing attention to these passages, I do not intend to challenge the claim that the sort of epistemic constraint which Dummett and the other neo-verificationists seek to impose on truth can be legitimately expressed through a statement like (K). However, Dummett's understanding of this principle is clearly informed by his other views about the role which the notion of proof should be assigned within a theory of meaning. These, in turn, are informed by the view that such a theory should be modeled on the notion of truth employed within intuitionistic logic and mathematics. And for this reason, there seems to be little motivation for thinking that a principle like (K) involving the unexplicated notion of knowability ought to be taken as the canonical means of expressing how an epistemic constraint on truth ought to be formulated.

If we look more broadly at the writings of the other neo-verificationists, it is easy to find evidence not only for this conclusion, but also for the view that proofs are taken to play the epistemic role of verifications within a theory of meaning. Both tendencies are clearly evident in the following passage from Martin-Löf:

A is true is taken to mean that there exists a proof of *A*, a proof which need not necessarily be direct or canonical. The term proof is of course synonymous with verification here. This definition of

the notion of truth of a proposition reduces it to two notions, namely the notion of proof or verification and the notion of existence, and it is because of this that it is very natural to use the term verificationism in connection with the theory of meaning for intuitionistic logic: the term verificationism is used to stress the fact that the notion of truth is not taken as a primitive notion, like a truth conditional theory of meaning, but is rather defined in terms of an underlying notion of verification by the principle that A is true if there exists a proof of A . ([69], p.191)

Although the views of Prawitz and Martin-Löf differ in many respects, Prawitz is also quite explicit in his endorsement of the view that truth should be defined in terms of proof existence. The following statement is characteristic:

Once we accept the notion of provability as legitimate, it is hardly controversial within verificationism that the truth of a proposition is to be identified with provability or existence of proofs. ... [P]roofs as here understood are something that in principle can be known by us, and hence there is no talk about in principle unknowable proofs. ([86], p. 48)

In order to appreciate the relationship between a statement such as this and Dummett's (K), a bit more must also be said about the difference between how Prawitz and Martin-Löf view the notion of proof itself. For Martin-Löf it is judgements which are the fundamental epistemic concept, while proofs derive their epistemic significance only from their relation to judgements. For Prawitz, on the other hand, proofs themselves are epistemic in nature in the sense that they are intrinsically the sorts of entities which can potentially be grasped by us – a point which he goes out of his way to emphasize in this passage.

The foregoing passages are typical of many others in the writings of the neo-verificationists in which the notion of truth is analyzed in terms of proof existence. They illustrate not only that a statement with the form (KP3) serves as a natural formalization of their views, but also that they are in agreement that such a principle should be taken as having the same general conceptual significance as (KP). But it should also be noted that from a purely logical perspective (KP) and (KP3) appear quite different. For while (KP) uses the combination of propositional operators “possibly known” to express knowability, (KP3) achieves this effect by using a single quantifier over proofs. Accepting a principle of this nature would generally be regarded as coming with a commitment to regarding proofs as objects. Such a commit-

ment also appears to be implicit in (NonOm2) which, in addition to quantifying over proofs, also presumes that it is legitimate to predicate of them the property of having been constructed. Although (KP4) and (NomOm3) represent natural formalizations of these statements in a formal language with proof quantifiers, adopting these statements as expressions of (KP3) and (NonOm2) only intensifies the need for adopting an objectual interpretation of proof quantification, since these latter statements employ the apparatus of predication theory to proof variables.

Given that we have already seen that the roots of neo-verificationism lie in intuitionistic mathematics, we must also confront the following questions in the course of justifying the adoption of (KP3) and (NonOm2):

- 1) Is it consistent with the general tenets of intuitionistic logic and mathematics to regard proofs as objects in a manner which legitimates the use of proof quantifiers?
- 2) Is it similarly admissible to distinguish between those proofs which have actually been constructed and those which have not been constructed and thus exist “merely potentially”?

These questions touch on some of the most difficult and controversial issues concerning the interpretation of intuitionism. As such, I will not attempt to provide satisfactory general answers here. What I will now attempt to do,

however, is to briefly describe how the views of the neo-verificationists on these matters bear on the legitimacy of adopting (KP3) and (NonOm2) in order to reason about the Knowability Paradox.

Modern expositions of the BHK interpretation often implicitly treat proofs as objects at least to the extent that they introduce schematic terms to denote proofs and operations on them. Beyond this, however, the “proofs as objects” view is explicitly adopted within several of the systems which attempt to formalize Heyting’s interpretation, perhaps most notably Kreisel’s *theory of constructions* [56]. This system corresponds to a formal calculus in which propositions, proofs and operations on proofs (i.e. constructions) are each assigned a syntactic representation. And although Kreisel claims this system to be “logic- and type-free”, not only the language of this theory but also its subsequent use by other theorists are all suggestive of the fact that it is indeed committed to regarding proofs as abstract, structured objects.¹¹

Dummett’s views on the objectual status of proofs are less clear. This

¹¹Such a conclusion is, for instance, already suggested by Kreisel’s use of a formal predicate $\Pi(c, A)$ to which he assigned the interpretation “ c is a proof of A .” Subsequent work on the theory of constructions by Goodman [40] led to the recognition that in order to ensure the consistency of Kreisel’s theory, it is necessary to regard constructions (which are identified with in proofs in the theory) as inhabiting stratified domains, the lowest of which consists of the rules of the original system and operations on them while higher levels are formed by reflecting on lower ones, conceived as completed domains. This picture is again suggestive of the “proofs as objects” view. Sundholm [100] arrives at the same conclusion on the basis of more general considerations about the role which constructions play within intuitionism.

is in part due to the fact that his discussions often hew more closely to the historical consideration of intuitionism than do those of Prawitz and Martin-Löf. However, many of the pronouncements he has made on other issues appear to be motivated by a fear that regarding proofs as abstract entities will entail that truth is bivalent, an outcome which he has repeatedly argued would signify a collapse into realism. One of my goals below will be to show that such a fear is ill-founded. However, we must first attempt to understand the basis of his worries. And in order to do this, it will be useful to turn our attention to the second question posed above – i.e. does it make sense to distinguish actual from potential proofs?

Prawitz and Martin-Löf have discussed the need for adopting such a view on a variety of occasions. Both theorists generally invoke the notion of potential proof in the context of arguing that mathematical truth is tenseless – i.e. that mathematical statements have truth values independently of our having proved them and thus do not “become” true only when then have been derived. Prawitz has put this point as follows:

Intuitionistic philosophers sometimes use true as synonymous with truth as known, but this is clearly a strange and unfortunate use. We need a notion of truth where, without falling into absurdities, we may say, e.g., that there are many truths that are not known

today. ([83], p. 8)

The necessity for distinguishing between actual and potential proofs comes from the need to reconcile this view with the position that the truth of a statement implies its provability. Prawitz describes how this can be accomplished as follows:

[I]t is here convenient to distinguish actual and potential existence ... We can then say that the correctness of an assertion requires the actual existence of a proof, while the truth of the asserted proposition requires only (and is identical with) the potential existence of a proof of the proposition.¹²([86], p. 48)

The picture which emerges from these passages is one according to which proofs are conceived as being brought into actual existence by the activity of mathematical agents, while at the same time acknowledging that such entities exist independently of their activities (i.e. in some “potential” sense). But this characterization leaves open the exact manner in which the distinction

¹²Several similar passages may be found in Martin-Löf – e.g.:

That proposition A is *actually true* means that A has been proved, that is that a proof of A has been actually constructed, which we can also express by saying that A is known to be true. ... [T]o say that A is *potentially true* is to say that A can be proved, that is that a proof of A can be constructed, which is the same as to say, in usual terminology, simply that A is true. ([68], p.141)

between actuality and potentiality should be understood. The problem comes out most clearly when we consider statements of the form “a proof of φ has not actually been constructed” which are relevant to the interpretation we wish to assign to (NonOm). For note that such a statement appears to be contrastable both with “a proof of φ *might* be constructed” and also “a proof of φ *will* be constructed.”¹³ I will propose a tentative means of making sense of the neo-verificationists’ use of these terms in Section 4.1 below.

One is also naturally led to ask whether the picture just described forces us to adopt the view that there is a fixed class encompassing all potential proofs which must be thought of as a “completed totality” – i.e. a determinate (and presumably) infinite class of (presumably) abstract objects. Such a conclusion might appear counter to the spirit of verificationism because of its apparent resemblance to the platonistic view that, e.g., the natural numbers form a set of abstract objects independently of our having constructed (representations of) them. If we additionally adopt the view that the falsity of a statement is equivalent to the non-existence of a proof, then it appears to follow from this view that all statements are determinately either true or false. Since such a conclusion could then be used to justify the law of the

¹³We may additionally note in this regard that Prawitz [84] often uses expression like “tenseless” to qualify what he intends by the notion of “potential existence.” It thus follows that if actual existence contrasts with potential existence, then to say that a proof exists not just potentially but actually as well, is presumably to say that it exists in time.

excluded middle for arbitrary statements, it will presumably be unacceptable to the neo-verificationists for, as we have seen, they are all proponents of intuitionistic logic.

Dummett has voiced both of these concerns. On the one hand:

We can introduce such a notion only by appeal to some platonistic conception of proofs as existing independently of our knowledge, that is, as abstract objects not brought into being by our thought. But if we admit such a conception of proofs, we can have no objection to a parallel conception of mathematical objects such as natural numbers, real numbers, metric spaces, etc. ([19], p. 258)

And on the other:

There is a well-known difficulty about thinking of mathematical proofs - and, equally, of verifications of empirical statements - as existing independently of our hitting on them, which insisting that there are proofs we are capable of grasping or of giving fails to resolve. Namely, it is hard to see how the equation of the falsity of a statement (the truth of its negation) with the non-existence of a proof or verification can be resisted: but, then, it is equally hard

to see how, on this conception of the existence of proofs, we can resist supposing that a proof of a given statement determinately either exists or fails to exist. We shall then have driven ourselves into a realist position, with a justification of bivalence. ([20], p. 285)

In reply to Dummett's concerns about non-constructed proof, it seems possible to cite the following argument based in the notion of decidability. Within intuitionism, a statement ψ is traditionally defined to be decidable if $\vdash_I \psi \vee \neg\psi$ where $\vdash_I$ denotes provability in an appropriate intuitionistic formal system (e.g. IQC or HA). This notion can be extended to an arbitrary predicate $P(x)$ over a domain D which is also decidable (i.e. one for which $\vdash_I \forall x(x \in D \vee x \notin D)$). In particular, we say that $P(x)$ is decidable just in case for all $a \in D$, $\vdash_I P(\bar{a}) \vee \neg P(\bar{a})$ where $\bar{a}$ is a name for a . The paradigmatic example of a decidable domain is the natural numbers. Paradigmatic examples of decidable predicates include *x is greater than 17*, *Even(x)* and *Prime(x)*. For when the definition of each of these predicates is written out as an open sentence $P(x)$ in the language of HA, it may be shown that $\text{HA} \vdash P(\bar{n}) \vee \neg P(\bar{n})$ for each natural number n .

Suppose we fix the example $P(x) = \text{Prime}(x)$. We then have

(4.4) For every natural number n , $\text{HA} \vdash \text{Prime}(\bar{n}) \vee \neg \text{Prime}(\bar{n})$.

Since HA satisfies the disjunction property, we additionally have

(4.5) For every natural number n , $\text{HA} \vdash \text{Prime}(\bar{n})$ or $\text{HA} \vdash \neg \text{Prime}(\bar{n})$.

Derivations within HA are presumably constructive proofs, *par excellence*.

And thus when formulated in the informal language of (KP3), (4.4) implies

(4.6) For every natural number n , there is a proof of $\text{Prime}(\bar{n})$ or there is a proof of $\neg \text{Prime}(\bar{n})$.¹⁴

Discussion of decidable predicates occur throughout standard expositions of intuitionism (notably including Dummett [17]). And hence there seems to be no reason to think that either (4.4) or (4.5) would be regarded as contentious by even orthodox intuitionists. Note, however, that the following statement also seems uncontentious

(4.7) There exists n such that we do not know either $\text{Prime}(\bar{n})$ or $\neg \text{Prime}(\bar{n})$.

It is notable that the existential quantifier in (4.7) can be read constructively in the sense that we can, at any given time, find an explicit value for n which makes the matrix of this statement true – e.g. as of June 2008, we can take $n = 2^{41,854,567} - 1$. Note, however, that it follows from this that we

cannot have constructed a proof of either $Prime(\bar{n})$ or $\neg Prime(\bar{n})$. For if we had, then surely we would know one statement or the other. And thus it appears to follow from (4.6) that even the most orthodox of intuitionists must acknowledge the existence of non-constructed or “merely potential” proofs.

The foregoing considerations suggest that the distinction between actual and potential proofs is not only intuitionistically legitimate but necessary in order to account for obvious facts about our epistemic situation within mathematics. We may additionally note, however, that (4.5) also appears to commit us to the existence of infinitely many potential proofs. Since at any given time, the sum of our mathematical activity can presumably have led to the construction of only a finite number of proofs, this suggests that we must accept the existence of proofs which have not been constructed. Note, however, that (4.5) only commits us to the existence of infinitely many proofs in the sense that it ensures that for every natural number there is a proof of $Prime(\bar{n})$ or $\neg Prime(\bar{n})$. There has generally been consensus among intuitionists that the natural numbers can legitimately be treated as a *potential infinity* – i.e. a class of entities given by a construction which can be indefinitely repeated. What follows from (4.6) is that for each n ,

¹⁴In fact, (4.6) can be proven internally in HA – i.e. $HA \vdash Bew_{HA}(\ulcorner \forall x [Bew_{HA}(\ulcorner Prime(x) \urcorner) \vee Bew_{HA}(\ulcorner \neg Prime(x) \urcorner)] \urcorner)$. Such a statement thus appears to be an undeniable consequence of principles universally accepted within intuitionism.

there exists a uniform method of constructing a proof $t(n)$ whose conclusion is either $Prime(\bar{n})$ or $\neg Prime(\bar{n})$. But this appears only to commit us to regarding potential proofs as corresponding to a potential infinity.

I will take the foregoing considerations to provide intuitionistic *bona fides* not only for distinguishing between actual and potential proofs but also for acknowledging that the latter category must be inhabited – i.e. that there exist potential proofs which have not been constructed. On this basis, the example of pairs of statements like $Prime(\bar{n})$, $\neg Prime(\bar{n})$ may additionally be used to make a related point which is of direct relevance to the Knowability Paradox. For as I have argued, we know (and in fact can prove – cf. note 14) that one or the other of these statements must be provable, despite the fact that neither has been proven. As such, it follows that there is no reason why the neo-verificationists should resist the conclusion that one of these statements is *true* (although of course we do not currently know which). If we now take χ to be whichever one of $Prime(\bar{n})$, $\neg Prime(\bar{n})$ is provable (and hence true), it follows that neo-verificationists should have no difficulty in accepting the conjunction

(4.8) χ and no proof of χ has been constructed.¹⁵

¹⁵It might be objected here that since no proof of either $Prime(\bar{n})$ or $\neg Prime(\bar{n})$ has been constructed, an intuitionist will not be able to assert a concrete instance of (NonOm3)

But this, of course, is just an instance of (NonOm2). As we observed in Section 1, one might naturally think that there are manifest reasons to be suspicious of such a statement from the standpoint of intuitionism. But it now appears that the intuitionistic admissibility of statements like (4.8) is a straightforward consequence of the traditional understanding of decidable predicates coupled with the observation that we can only have constructed a finite number of proofs at any given time.

4.3 Reconstruction and consistency

In the previous section I argued that (KP3) and (NonOm2) represent plausible reconstructions of how the neo-verificationists are likely to understand the epistemic principles on which the Knowability Paradox is premised.

It remains to be shown that these principles do not lead to either a contradiction or other counter-intuitive consequences. I will pursue this goal in

but only the intuitionistically weaker statement

(NonOm2d) $\psi \vee \neg\psi$ and no proof of either ψ or $\neg\psi$ has been constructed.

I have presented a case for the intuitionistic legitimacy of (NonOm2) itself mainly to preserve parity with the original exposition of the Paradox. However the current observation might be taken to weaken the claim that intuitionists are committed to accepting (NonOm1) in the original derivation of the Paradox and thereby also the conclusion (Om). But it may be shown in a manner similar to that employed by Percival [79] that a version of the Paradox survives even if we weaken (NonOm1) to

(NonOm1d) For some ψ , $(\psi \vee \neg\psi) \wedge (\neg K\psi \wedge \neg K\neg\psi)$.

This observation should be compared with that given in note 17 below which formalizes the fact that the existence of non-constructed proofs follows from (KP4) by intuitionistically valid reasoning even if we assume (NonOm2d) instead of (NonOm2).

three stages. First I will present the formal specification of the system QLPE in which the proposed formalizations of these statements (i.e. (KP4) and (NonOm3)) are expressed. Second, I will argue that no statement analogous to (Om) can be derived in QLPE from (KP4) and (NonOm3) in a manner which mimics Fitch's original derivation. And third I will then demonstrate that these principles do not conflict in any other way by presenting a semantics for QLPE relative to which they can be shown to be formally consistent.

4.3.1 QLPE and the reconstruction of the Fitch derivation

As mentioned above, QLPE represents only a very minor modification of the form of QLP presented in Chapter 2. The only change to the language of this system is the addition of the unary predicate $E(x)$, which results in a language which I will call $\mathcal{L}_E$. The class of proof terms of QLPE is thus identical that of QLP and the formal definition of its class of formulas of $\mathcal{L}_E$ is as follows:

Definition 4.3.1. For t a proof term, the class of formulas of $\mathcal{L}_E$ is specified by the grammar

$$\varphi := P_i \mid E(t) \mid \varphi_1 \wedge \varphi_2 \mid \varphi_1 \vee \varphi_2 \mid \varphi_1 \rightarrow \varphi_2 \mid \neg \varphi_1 \mid t : \varphi_1 \mid (\forall x)\varphi_1 \mid (\exists x)\varphi_1$$

The free variables of $\varphi(x)$ are defined in the standard manner of first-order

logic. The class of sentences of $\mathcal{L}_E$ is accordingly defined as those formulas which contain no free variables.

We must now ask how we ought to axiomatize reasoning in the language $\mathcal{L}_E$ relative to the intended interpretation of $E(x)$ and the proof quantifiers adopted above. The most pressing question is obviously what, if any, new axioms ought to be adopted involving the predicate $E(x)$. It might seem that we should adopt axioms require express that the extension of $E(x)$ is closed in various ways . Many such closure properties are easily expressed in $\mathcal{L}_E$ – e.g.

$$(EApp) \quad (\forall x)(\forall y)[(E(x) \wedge E(y)) \rightarrow E(x \cdot y)]$$

$$(EDown) \quad (\forall x)(\forall y)[E(x \cdot y) \rightarrow (E(x) \wedge E(y))]$$

$$(EP) \quad E(p(x_1, \dots, x_n)) \text{ for } p(x_1, \dots, x_n) \text{ a primitive proof term}$$

Although a case can be made in favor of these and a variety of similar closure principles for $E(x)$, the strategy I will adopt here is to avoid imposing any restrictions on this predicate whatsoever. It follows from this that the axioms and rules we adopt for QLPE are identical with those of QLP.

This decision is in keeping with the view expressed above that whether a particular proof is constructed is a contingent matter and thus not obviously

amenable to logical characterization. Note, for instance, that if we accord EApp the status of an axiom, then it will then follow that $E(x)$ must be closed under reasoning mediated by the QLP application axiom. As such, if we also adopt a principle like EP which requires that all primitive proof terms are constructed, it will immediately follow that $E(x)$ must hold of infinitely many proof terms. But this is presumably out of keeping with the desire of the neo-verificationist to avoid a commitment to regarding the class of proofs as actually infinite. Although the policy of not assuming any axioms involving $E(x)$ may seem like an overly extreme reaction to this problem, we will see below that closure principles in fact have only a minor bearing on the status of the Knowability Paradox.

We may now turn to the task of reconstructing the reasoning of the Paradox using the $\mathcal{L}_E$ formalizations of (KP3) as (KP4) and (NonOm2) as (NonOm3). I reproduce these principles here for reference:

(KP4) For all φ , $\varphi \rightarrow (\exists x)x : \varphi$.

(NonOm3) For some ψ , $\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]$.

In order to attempt to recreate the reasoning of the paradox, we wish to see what happens when we adjoin these statements to QLPE. To facilitate this, let $\vdash_{\text{QLPE}+\text{KP4}}$ denote derivability in the system consisting of QLPE together

with all instances of (KP4) and let NonOm3_χ denote the single statement formed by taking $\psi \equiv \chi$ in (NonOm3) .¹⁶

Proposition 4.3.1. For any $\psi \in \text{Sent}_E$, $\text{NonOm3}_\psi \vdash_{\text{QLPE+KP4}} (\exists x)\neg E(x)$.

Proof. Writing $\vdash$ for $\vdash_{\text{QLPE+KP4}}$ we may reason as follows:

- | | |
|--|-------------------------|
| 1) $\text{NonOm3}_\psi \vdash \psi$ | propositional logic |
| 2) $\text{NonOm3}_\psi \vdash (\forall x)[E(x) \rightarrow \neg x : \psi]$ | propositional logic |
| 3) $\text{NonOm3}_\psi \vdash \psi \rightarrow (\exists x)x : \psi$ | KP4 |
| 4) $\text{NonOm3}_\psi \vdash (\exists x)x : \psi$ | 1), 3) |
| 5) $\text{NonOm3}_\psi \vdash (\forall x)[E(x) \rightarrow \neg x : \psi] \rightarrow$
$((\exists x)x : \psi \rightarrow (\exists x)\neg E(x))$ | QLP1, QLP2, QLP4, RQLP1 |
| 6) $\text{NonOm3}_\psi \vdash (\exists x)x : \psi \rightarrow (\exists x)\neg E(x)$ | 2), 5) |
| 7) $\text{NonOm3}_\psi \vdash (\exists x)\neg E(x)$ | 4), 6) |

□

This result demonstrates that when formalized in $\mathcal{L}_E$, the premises of the Knowability Paradox lead directly to the conclusion that there are non-constructed proofs. This is in line with the conclusion reached at in the

¹⁶For purposes of definiteness, we do not treat (KP4) as an axiom and hence RPL2 does not apply to it. Nothing in the sequel depends on this.

previous section on the basis of decidable predicates and has nothing directly to do with the Fitch derivation itself.¹⁷ But one is also naturally drawn to ask what happens when we attempt to reconstruct the original Fitch derivation on the basis of these premises. As I will now attempt to demonstrate, such an attempt does not lead to a reiteration of the Paradox. It does, however, provide another way of demonstrating how (NomOm3) entails the existence of non-constructed proofs which is of some interest in itself.

In order to reconstruct the reasoning of the Fitch derivation, I will work in the system $\text{QLPE} + \text{KP4}$ and this time treat NonOm3_ψ as a reductio assumption. In parallel to the original derivation, we endeavor to derive

(Om3) For all $\varphi \in \mathcal{L}_E$, $\varphi \rightarrow (\exists x)[E(x) \wedge x : \varphi]$.

by taking NonOm3_ψ as the substitution instance for φ in (KP4). This gives

us

1') $\text{NonOm3}_\psi \vdash \psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]$

¹⁷It may also be noted that $(\exists x)\neg E(x)$ is also derivable in $\text{QLPE} + \text{KP4}$ on the basis of somewhat weaker premises than those employed in Proposition 4.3.1. In particular, suppose on the basis of the discussion in note 15 about the assertability of a particular instance of (NonOm3) that a neo-verificationist is only willing to accept

(NonOm3d) $(\psi \vee \neg\psi) \wedge (\forall x)[E(x) \rightarrow (\neg x : \psi \wedge \neg x : \neg\psi)]$

It follows by essentially the same reasoning employed in Proposition 4.3.1, plus an intuitionistically valid proof by cases, that for any ψ , $\text{NonOm3d}_\psi \vdash_{\text{QLPE}+\text{KP4}} (\exists x)\neg E(x)$. We take this to demonstrate that a commitment to the existence of non-constructed proofs is already implicit in (KP4) even when taken in conjunction with very weak formulations of non-omniscience.

$$2') \text{ NonOm}3_\psi \vdash (\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]) \rightarrow$$

$$(\exists y)y : [\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]] \quad \text{KP4}$$

$$3') \text{ NonOm}3_\psi \vdash (\exists y)y : [\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]] \quad 1'), 2')$$

In the original derivation at line 4), the statement $K(\psi \wedge \neg K\psi)$ is assumed as another reductio premise. If we continue to use the analysis of knowledge given by (4.2b), then this statement should go over into $\mathcal{L}_E$ as

$$(4.9) \quad (\exists z)[E(z) \wedge z : (\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi])]$$

In order to mimic the effect of assuming this statement as an open assumption in a natural deduction proof, it is most convenient to introduce a free-variable instance of this statement $E(z) \wedge z : (\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]) \equiv \Theta$ as an assumption. On this basis, we may continue as follows:

$$4') \quad \Theta \vdash z : [\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]] \quad \text{propositional logic}$$

$$5') \quad \Theta \vdash E(z) \quad \text{propositional logic}$$

$$6') \quad \Theta \vdash a_1 : ([\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]] \rightarrow \psi) \quad \text{RQLP3}$$

$$7') \quad \Theta \vdash a_2 : ([\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]] \rightarrow$$

$$\forall x[E(x) \rightarrow \neg x : \psi]) \quad \text{RQLP3}^{18}$$

$$8') \quad \Theta \vdash a_1 \cdot z : \psi \quad \text{LP1 5'), 6')}$$

$$9') \quad \Theta \vdash a_2 \cdot z : (\forall x)[E(x) \rightarrow \neg x : \psi] \quad \text{LP1 5'), 7')}$$

$$10') \quad \Theta \vdash a_2 \cdot z : (\forall x)[E(x) \rightarrow \neg x : \psi] \rightarrow$$

	$(\forall x)[E(x) \rightarrow \neg x : \psi]$	LP3
11')	$\Theta \vdash (\forall x)[E(x) \rightarrow \neg x : \psi]$	9'), 10')
12')	$\Theta \vdash E(a_1 \cdot z) \rightarrow \neg a_1 \cdot z : \psi$	QLP 11')
13')	$\Theta \vdash \neg E(a_1 \cdot z)$	propositional logic

Steps 4')-13') can be taken to parallel steps 4)-9) in the original derivation – i.e. the hypothetical subproof with premise $K(\psi \wedge \neg K\psi)$. At the end of this subproof we were originally able to derive $\neg K(\psi \wedge \neg K\psi)$ based on the conflict between step 7) (i.e. $K\psi$) and step 9) (i.e. $\neg K\psi$). In the reconstructed derivation, however, we derive only $E(z)$ and $\neg E(a_1 \cdot z)$. These statements respectively express that the verification denoted by z has been constructed and that the verification denoted by $a_1 \cdot z$ has not. Note that this is not a formal contradiction relative to the axioms we have presented for QLPE. For pursuant to the discussion above, it is at least consistent to think that we may have constructed the verification denoted by z but have not performed the additional step of applying a_1 to it so as to construct the verification denoted by $a_1 \cdot z$.

¹⁸Steps 5') and 6') respectively follow by applying axiom necessitation to the tautologies $(P \wedge Q) \rightarrow P$ and $(P \wedge Q) \rightarrow Q$ with $P \equiv \psi$ and $Q \equiv (\forall x)[E(x) \rightarrow \neg x : \psi]$. Here we take a_1, a_2 such that $(P \wedge Q) \rightarrow P \in \mathcal{P}(a_1)$ and $(P \wedge Q) \rightarrow Q \in \mathcal{P}(a_2)$ where $\mathcal{P}$ is some axiomatically appropriate constant specification.

One might take this to imply that the original derivation of the Paradox breaks down in QLPE only because we have not elected to impose closure axioms on the extension of $E(x)$ such as EApp and EAx. In order to see that this is not the case, we may simply assume these principles and continue to reason in the system QLPE + KP4 + EAx + EApp.

15')	$\Theta \vdash E(a_1)$	EAx
16')	$\Theta \vdash (E(a_1) \wedge E(z)) \rightarrow E(a_1 \cdot z)$	RQLP1, EApp
17')	$\Theta \vdash E(a_1 \cdot z)$	5'), 15'), 16')
18')	$\Theta \vdash \perp$	13'), 17')
19')	$\vdash \neg(E(z) \wedge z : (\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]))$	18'), deduction theorem
20')	$\vdash \neg(\exists z)(E(z) \wedge z : (\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]))$	19'), RQLP2

The derivation 4')-20') can be taken to structurally parallel steps 4)-10) in the Fitch derivation. Step 10) in the original derivation can be taken to express that it cannot be known that $\psi \wedge \neg K\psi$. But since this conclusion is provable from no premises, it follows by the modal necessitation rule that 11) $\Box \neg K(\psi \wedge \neg K\psi)$, which in turn can be taken to express that $\psi \wedge \neg K\psi$ is an *unknowable* truth. A parallel claim in the current setting would be that $\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]$ is an *unprovable* truth, contra 3'). But note

that 20') expresses merely that no proof of this statement can fall under the extension of E . Relative to the interpretation of $\mathcal{L}_E$ introduced above, this does not entail the claim that this statement is unprovable, but merely that no proof of this statement has already been constructed.

Although this conclusion does not stand in formal conflict with 3'), one might at first think there was still a sort of conceptual tension between 20') and 3'). For while on the one hand 3') reports that there is a proof z of $\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]$, the interpretation just provided for 20') appears to suggest that this proof can never be constructed. One might thus be led to believe that z is a sort of “ideal” proof which exists in principle but could never be actualized in practice. For recall that the fundamental motivation behind the neo-verificationist’s adoption of a principle such as (KP) is to rule out the existence of unverifiable truths. As such, the conclusion that there are true statements which are verifiable only by non-constructible proofs might be taken to show that the formalization of (KP) by (KP4) does not adequately capture the intentions of these theorists in advocating an epistemic constraint on truth.

In order to see that the tension just described is merely apparent, some additional consideration must be paid to the role played by the predicate $E(x)$ within QLPE. As stated above, this predicate is intended to mark a

distinction between actual and potential proofs which I have argued is both acknowledged in the writings of the neo-verificationists and required in virtue of standard assumptions about the status of decidable predicates within intuitionism. The question remains as to how this distinction can be made precise with respect to a description of our actual deductive activities. I will consider this question in more detail in Section 4. For now, however, it will suffice to note that it is at least consistent with the writing of the neo-verificationists to think of the quantifiers of QLPE as ranging over all verifications which are constructible in the sense of corresponding to the denotations of proof terms t . Note, however, that this class is inductively generated and hence infinite. Thus at any given stage in our deductive activity, we will only have constructed finitely many of its members. The predicate $E(x)$ is intended to hold of those proof terms denoting verifications we have constructed at or prior to the current stage.¹⁹

¹⁹The relevant contrast between constructible and constructed proofs can also be characterized by likening the quantifier $(\exists x)$ to a *possibilist* quantifier over proofs – i.e. one which ranges not only over proofs which are actual (in the sense of having been constructed) but over ones which are merely possible (in the sense of being constructible in principle). This helps to also explain why $\mathcal{L}_E$ sentences of the form $(\exists x)x : \varphi$ can be taken to express the *knowability* of φ . Note that if we were to adopt an *actualist* interpretation of proof quantification – i.e. one in which proof quantifiers ranged only over proofs which had actually been constructed – we would be forced to introduce traditional modal operators to express statements about provability in principle. (KP3) and (NonOm2) would then be respectively formalized as $(\text{KP}_m) \varphi \rightarrow \Diamond(\exists x)x : \varphi$ and $(\text{NonOm}_m) \psi \wedge \neg(\exists x)x : \psi$. However, it can readily be verified that these principles lead to a contradiction via reasoning which parallels the Fitch derivation. As such, a possibilist interpretation of the QLPE quantifiers appears to be preferred both because it appears to conform better with the use of proof

If we apply this interpretation to 20'), this statement no longer has the appearance of expressing that no proof of $\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]$ is constructible in principle. This proposition would be expressed in QLPE by a statement with an unrestricted proof quantifier of the form

$$(4.10) \quad \neg(\exists y)y : (\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi])$$

This statement expresses that there is no proof (and therefore no constructible proof) of $\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]$ and as such obviously stands in contradiction to 3'). But, 20') differs from (4.10) in that the former begins with a quantifier restricted by $E(x)$. It consequently expresses not that there is no proof of $\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]$ (or that it is impossible to construct such a proof), but merely that we cannot currently have constructed a proof of a statement of this form.

The provability of 20') in QLPE thus does not reflect that $\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]$ possesses some mysterious epistemic property. Rather it codifies the more mundane observation that were we to have constructed a proof of such a statement at the present stage in our deductive activity, we would thereby have refuted its second conjunct. For suppose that z verifies $\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]$ and also that $E(z)$. Then it follows both that $a_1 \cdot z$

quantification in the writings of the neo-verificationists and also because it allows us to avoid becoming trapped in the original reasoning of the Paradox.

verifies ψ and also that $(\forall x)[E(x) \rightarrow \neg x : \psi]$ must be true. But this latter statement expresses that no verification of ψ has been constructed, meaning in particular that $\neg E(a_1 \cdot z)$. If we again assume the closure EP and EApp on $E(x)$, it follows that z itself cannot fall under E on pain of contradicting this conclusion.

Hand [46], [47] has called attention to essentially the same phenomena in the context of the traditional presentation of the Knowability Paradox. He suggests that the situation just described can be taken to reflect a genuine limitation on our ability to construct proofs of statements which describe the current extent of our knowledge. In particular, he suggests that we cannot, as a matter of practical fact, verify a conjunction like $\psi \wedge \neg K\psi$ one of whose constituents expresses that the other has not been verified. For in the course of so doing we would falsify the second conjunct. Hand maintains on behalf of the neo-verificationists that such statements might still be verifiable in the sense that there exists a constructive procedure by which they could be shown to be true. But for the operational reason just described, such a procedure could be carried out in practice.

The sort of resolution to the Paradox proposed here relies on a related distinction between constructed and constructible proofs. For although 20') reports that no verification of $\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi])$ can be constructed

now, this is not in formal conflict with 3') which states that such a proof exists. What 20') does entail, however, is that if $\text{NonOm}3_\psi$ is true for any ψ , then (KP4) entails that there must exist a non-constructed proof. Although this is essentially the same result as Proposition 4.3.1, 4')-13') show how the result can be reached in a manner which resembles the original reasoning of Fitch. This same finding can be presented more perspicuously by subsuming $\text{NonOm}3_\psi$ into the antecedent of a conditional to yield

Proposition 4.3.2. For any formula ψ of $\mathcal{L}_E$,

$$\vdash_{\text{QLPE}} (\forall z)(z : [\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]] \rightarrow \neg E(a_1 \cdot z))$$

On the basis of the foregoing discussion, it should now be clear that there is no sense in which the verification promised by 3') of $\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]$ must be viewed as a *non-constructible* proof. For note that there is nothing about 20') or Proposition 4.3.2 which precludes that a verification of $\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]$ might come to be constructed at some later point in our deductive practice than that which $E(x)$ describes. Put in terms of the original distinction drawn by the neo-verificationists, this is to say that while the proof claimed to exist by 3') must of necessity be merely potential, there is no in-principle reason why it cannot be made actual by further deductive effort.²⁰

²⁰This means of resolving the issue does, of course, depend on the assumption that $E(x)$

The foregoing considerations also appear to suggest that there is no straightforward way of continuing the above derivation in QLPE + (KP4) which eventuates in (Om3) via the reasoning of the Fitch derivation.²¹ However, in order to demonstrate this formally, it is not sufficient to merely note that there is no contradiction between 3') and 20'). Rather, we must somehow show that for at least certain choices of ψ , NonOm3 $_{\psi}$ is consistent with all instances of (KP4). I will now turn to this task by adapting the semantics presented for QLP in Chapter 2 to QLPE. This will allow us to both prove such a consistency result and also to frame more precisely the distinction between actual and potential proofs which we have been discussing.

does not vary in extension over time but always holds of just those verifications which have been constructed at or prior to the present moment. This might be achieved, for instance, by determining its extension indexically by referring to the current moment through the use of a word like 'now' in a manner which also allowed us to make subsequent reference to this time. I will return to question of whether such indexical devices are ultimately required in order to fix the extension of E in Section 4.1.

²¹One reason for this is that QLPE has no precise analogue to the modal necessitation rule by which the step from 10) to 11) is mediated in the Fitch derivation. It is possible to mimic this step indirectly by applying the Lifting Lemma to produce a proof term t such that $\vdash t : \neg(\exists z)(E(z) \wedge z : (\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]))$ and then existentially generalizing to obtain $\vdash (\exists u)u : \neg(\exists z)(E(z) \wedge z : (\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]))$. If provability is likened to a kind of necessity, then this statement can be taken to express the necessity of the non-existence of a *constructed* proof of $\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi]$. However, again this does not rule out the possibility that there might exist a non-constructed proof of this statement and therefore does not stand in conflict with 3').

4.3.2 The consistency of (KP4) and (NonOm3)

The foregoing discussion suggests that the consistency of (KP4) and (NonOm3) rest on the admissibility of a situation in which an instance of (NonOm3) is provable, but its proof has not been constructed. Turning this observation into a formal proof which ensures that no contradiction can be derived from (KP4) and (NonOm3) requires that we show the formal consistency of the set

$$\Gamma_\psi = \{\psi \wedge \forall x[E(x) \rightarrow \neg x : \psi]\} \cup \{\varphi \rightarrow (\exists x)x : \varphi \mid \varphi \text{ is a sentence of } \mathcal{L}_E\}$$

consisting of an appropriately chosen instance (NonOm3_ψ) of non-omniscience together with all instances of (KP4) by constructing a model of QLPE in which Γ is satisfied. This can be done in a straightforward manner by giving a semantics for QLPE and then showing that there is an appropriate model in which Γ_ψ is satisfied.

Note, however, that we face a choice in how to adapt the semantics developed for QLP in Chapter 2 to QLPE. The option I will adopt here is to simply adjoin to the definition of model a second domain function $\mathcal{D}_E$ which assigns to each world w a set of reasons (i.e. denotations of proof terms) $\mathcal{D}_E(w)$ which is a subset of the function $\mathcal{D}(w)$ which is used to interpret $E(x)$ at w . But note that if we do this, we are immediately drawn to ask how the sets

$\mathcal{D}_E(w_1)$ and $\mathcal{D}_E(w_2)$ are related to each other relative to accessibility in the model in question. For instance, ought we to assume that $\mathcal{D}_E(w)$ remains constant across worlds? that it expands monotonicity in the case that w_2 is accessible from world w_1 ? etc.

In attempting to answer such questions, it should be noted that the resolution to the Knowability Paradox which I have advocated above does not rely on modal notions in any essential way. In particular, although I have argued that the concept of knowability which figures in the appropriate understanding of (KP3) should be understood in terms of potential proof, I have argued that the latter concept should not be understood in the counterfactual sense of possibility. Rather I have argued that it ought to be understood in contradistinction to the notion of a proof which has actually been constructed. From this it follows that there is no apparent need to view a merely potential proof as something which does not exist in the actual world. It follows that the distinction between potential and actual proofs can be characterized non-modally by simply fixing a subset of the domain of each world which be understood as the class of proofs which have been constructed there. And for this reason, we make no requirement on the behavior of $\mathcal{D}_E(w)$ relative to the accessibility relation of a model. (In fact, we will see in a moment that it is possible to prove the consistency of Γ_ψ with respect to a one-world

model. So the question of the interpretation of the accessibility relation and its relationship to $E(x)$ in QLPE models drops out entirely.)

To this end, we define a QLPE model $\mathcal{M}$ to be a structure of the form $\mathcal{M} = \langle W, R, \mathcal{D}, \mathcal{D}_E, \mathcal{J}, \mathcal{E}, V \rangle$ such that $\mathcal{M} = \langle W, R, \mathcal{D}, \mathcal{J}, \mathcal{E}, V \rangle$ is a QLP model and $\mathcal{D}_E$ is a function from W to subsets of $\overline{\mathcal{D}}$ such that $\mathcal{D}_E(w) \subseteq \mathcal{D}(w)$ for all $w \in W$. All other definitions from Chapter 2 (e.g. of a formula being meaningful at a world, of a model meeting a primitive term specification, etc.) remain unchanged. The definition of truth in a model also remains unchanged from the QLP case with the exception that we now have the clause

- $\mathcal{M}, w \Vdash_v E(t)$ iff $t^v \in \mathcal{D}_E(w)$

Since the axiomatization of QLPE is identical to that of QLP, we immediately have the following:

Theorem 4.3.1. (Soundness) For all axiomatically appropriate primitive term specifications $\mathcal{P}$, if $\Gamma \vdash_{\mathcal{P}} \varphi$ then $\Gamma \Vdash^{\mathcal{P}} \varphi$.

As we will see in, the Soundness Theorem is already sufficient to prove the consistency of Γ_{ψ} with respect to appropriate ψ . However, it is also worth noting that since the axioms of QLPE make no mention of the $E(x)$ predicate, the completeness of QLPE with respect to the semantics just presented is also

an easy corollary of the completeness proof for QLP. In particular, we may modify the definition of the canonical $\mathcal{M}^*$ model of QLP to obtain a canonical model for $\mathcal{M}_E^*$ for QLPE by defining $\mathcal{D}_E(\Gamma)$ to be the set of pseudo-closed terms t such that $\Gamma \vdash E(t)$ where Γ is a world-like maximally consistent set of QLPE formulas. We then have

Theorem 4.3.2. Suppose that $\mathcal{M} \Vdash^{\mathcal{P}} \varphi$ for all QLPE models $\mathcal{M}$, then $\Gamma \vdash_{\mathcal{P}} \varphi$.

We now turn to question of proving the consistency of Γ_{ψ} relative to the semantics just presented. The first thing to note is that we cannot hope to prove the consistency of this set for *all* choices of ψ for the trivial reason that ψ may imply the negation of an instance of (KP4). However, we do get the following:

Proposition 4.3.3. Let Q be an arbitrary propositional letter. Then $\Gamma_Q \not\vdash \perp$.

Proof. By the Soundness Theorem, it suffices to construct a QLPE model $\mathcal{M}$ such that for some world w in $\mathcal{M}$ and valuation v , $\mathcal{M}, w \Vdash_v \Gamma_Q$. To this end, let $\mathcal{M}' = \langle W, R, \mathcal{D}', \mathcal{D}_E, \mathcal{J}, \mathcal{E}', V \rangle$ be arbitrary such that W contains the single world w accessible from itself, such that $w \in V(Q)$ and $\mathcal{D}_E(w) = \emptyset$. Clearly $\mathcal{M}, w \Vdash_v Q \wedge \forall x[E(x) \rightarrow \neg x : Q]$ for any v . Let r be a new reason not in $\mathcal{D}'(w)$ and let $\mathcal{D}(w) = \mathcal{D}'(w) \cup \{r\}$. Define $\mathcal{E}_0$ as follows:

i) $\mathcal{E}_0(w, s, v) = \mathcal{E}'(w, s, v)$ for all v and $s \neq r$

ii) $\mathcal{E}_0(w, r, v) = \{\varphi \mid \varphi \text{ is a sentence of } \mathcal{L}_E\}$

Let $\mathcal{E}$ be the result of closing $\mathcal{E}_0$ under conditions on evidence functions given in Chapter 2. Then $\mathcal{M} = \langle W, R, \mathcal{D}, \mathcal{D}_E, \mathcal{J}, \mathcal{E}, V \rangle$ is also QLPE model. We now claim $\mathcal{M}, w \Vdash_v \Gamma_Q$ for all v . Note that since we have changed neither V nor $\mathcal{D}_E$, $\mathcal{M}, w \Vdash_v Q \wedge \forall x[E(x) \rightarrow \neg x : Q]$ for any v . Now suppose $\mathcal{M}, w \Vdash_v \varphi$. Then $\mathcal{M}, w \Vdash_{v[x/r]} \varphi$ (since φ is a sentence and thus has no free variables). Note that $\varphi \in \mathcal{E}(r, w, v)$ and hence also $\mathcal{M}^* \Vdash_{v[x/r]} x : \varphi$ since w is accessible only to itself. Thus $\mathcal{M}, w \Vdash_v (\exists x)x : \varphi$ and hence $\mathcal{M}, w \Vdash_v \varphi \rightarrow (\exists x)x : \varphi$. Since φ was arbitrary, we thus have $\mathcal{M}, w \Vdash_v \Gamma_Q$. $\square$

It may be noted with respect to the foregoing proof that the model $\mathcal{M}$ constructed in order to show the consistency of (KP4) and (NonOm3) is such that $\mathcal{M}, w \Vdash_v \neg(\exists x)E(x)$ and hence has an empty domain of constructed proofs. It is thus natural to ask whether the consistency of Γ_ψ depends on our decision to not include closure principles for $E(x)$ among the axioms of QLPE itself. Some delicacy is required in order to answer this question, since we have seen it may not be clear what sort of statements should be counted as reasonable closure principles for constructed proofs. Call a proof term t *almost closed* if it contains no free variables except possibly in primitive proof terms and let

$\mathcal{C}$ be the set of such terms relative to some fixed primitive term specification $\mathcal{P}$ (as always we assume that $\mathcal{P}$ satisfies the free variable condition – i.e. if $\varphi(x_1, \dots, x_n) \in \mathcal{P}(a(x_1, \dots, x_n))$ then $\varphi(y_1, \dots, y_n) \in \mathcal{P}(a(y_1, \dots, y_n))$). A natural maximal condition is to assume that $E(t)$ holds of *all* terms in $\mathcal{C}$. And thus the axiom schema

(CP) For all $t \in \mathcal{C}$, $E(t)$.

expresses quite a strong closure principle relative to the considerations discussed above. (Note that it follows from the Lifting Lemma that if $\text{QLPE} \vdash \varphi$ and φ is closed then $\text{QLPE} \vdash t : \varphi$ for some $t \in \mathcal{C}$. It hence follows that if we assume $\text{QLPE} + (\text{CP}) \vdash (\exists x)[E(x) \wedge x : \varphi]$ whenever $\text{QLPE} \vdash \varphi$.)

In order to demonstrate that (CS) does not bring (NonOm3) and (KP4) into conflict, we may note the following:

Corollary 4.3.1. Let Q be an arbitrary propositional letter. Then $\Gamma_Q \cup \{E(t) \mid t \in \mathcal{C}\} \not\vdash \perp$.

Proof. By the Soundness Theorem, it again suffices to construct $\mathcal{M}, w$ and v such that $\mathcal{M}, w \Vdash_v \Gamma_Q \cup \{E(t) \mid t \in \mathcal{C}\}$. As in the canonical model construction for QLPE, let $p_0, p_1, \dots$ be an infinite set of parameters, $\mathcal{C}^*$ the set of pseudo-closed terms relative to these parameters and $\mathcal{P}^*$ the primitive specification derived from $\mathcal{P}$ by taking $\varphi(p_0, \dots, p_n) \in \mathcal{P}(a)$ iff $\varphi(x_1, \dots, x_n) \in \mathcal{P}(a)$.

We first define a structure $\mathcal{M}' = \langle W, R, \mathcal{D}', \mathcal{D}_E, \mathcal{J}, \mathcal{E}', V \rangle$ as follows:

- $W = \{w\}$, $R = \{\langle w, w \rangle\}$
- $\mathcal{D}'(w) = \mathcal{D}_E(w) = \mathcal{C}^*$
- $\mathcal{J}$ is defined in the manner of the interpretation of the canonical model of QLPE (i.e. $\mathcal{J}$ is the identity on primitive proof terms and terms of the form $!t$, $t_1 \cdot t_2$ and $t_1 + t_2$ and $(\langle t \forall x \rangle)^{\mathcal{J}} = \langle \lambda^* x. t \rangle$).
- $\varphi \in \mathcal{E}'(w, t^v, v)$ iff $\text{QLPE}(\mathcal{P}^*) \vdash (t : \varphi)\sigma_v$
- $V(Q) = \{w\}$ and $V(P) = \emptyset$ for all other propositional letters.

As it is straightforward to verify that $\mathcal{E}'$ satisfies the evidence closure condition, $\mathcal{M}'$ is hence a model of QLPE. Since $\text{QLPE}(\mathcal{P}^*) \not\vdash Q$, it hence follow that for all valuations v , $\mathcal{M}, w \not\models_v (t : Q)\sigma_v$ for any $t \in \mathcal{C}^*$. Hence since Q is true at w in $\mathcal{M}$, $\mathcal{M}', w \models_v Q \wedge \forall x[E(x) \rightarrow \neg x : Q]$. We made now proceed to expand $\mathcal{M}'$ to $\mathcal{M}$ in the same manner as Proposition 4.3.3 – i.e. by letting $\mathcal{D}(w) = \mathcal{D}'(w) \cup \{r\}$, adding φ to $\mathcal{E}'(\varphi, r, w)$ for all sentences φ and then defining $\mathcal{E}$ by closing under the evidence conditions. Via the same argument as above we obtain $\mathcal{M}, w \models_v \varphi \rightarrow (\exists x)x : \varphi$ for all φ and since we have changed neither V nor $\mathcal{D}_E(w)$, $\mathcal{M}', w \models_v Q \wedge \forall x[E(x) \rightarrow \neg x : Q]$ as well. Note finally that we also have $\mathcal{M} \models_v E(t)$ where v is the valuation

which replaces any free variable in x_i with the corresponding parameter p_i .

Hence $\mathcal{M}, w \Vdash_v \Gamma_Q \cup \{E(t) \mid t \in \mathcal{C}\}$. $\square$

The foregoing results show that regardless of whether we assume that all reasons denoted by almost closed terms are constructed, any statement of the form $Q \wedge (\forall x)[E(x) \rightarrow \neg x : Q]$ can consistently be conjoined with (KP4). This case appears to cover the sorts of (potential) examples of non-omniscience which are typically discussed in relation to the Knowability Paradox. For in attempting to illustrate the intuitive plausibility of the fact that there are true but unknown propositions, commentators have focused on contingent empirical statements which might express truths which no one ever bothers to verify – e.g.: “The comet contains pre-biotic molecules” [24], “The number of tennis balls in Timothy Williamson’s garden on 4 July 90 was even” [115], “The number of hairs on Wolfgang Künne’s head on 1 January 2007 was odd” [60]. Note that such statements neither possess logical structure which can be discerned in propositional languages like $\mathcal{L}_{LK}$ or $\mathcal{L}_E$, nor do they themselves contain epistemic operators. As such these situations fall naturally under the scope of Proposition 4.3.3.

4.4 Comparisons and contrasts

A useful taxonomy of strategies for resolving the Knowability Paradox is provided by Brogaard and Salerno [10], who distinguish between what they refer to as *logical revisionism*, *syntactic restrictionism* and *semantic revisionism*.²² The use of these expressions should be understood relative to the original derivation of the Paradox based on (KP1) and (NonOm1). Logical revisionists seek to resolve the Paradox by changing the underlying logic which is employed to reason about these principles in a manner which prevents the Fitch derivation from going through. Syntactic restrictionists similarly seek to block the derivation by limiting the class of allowable substitution instances in (KP1). Semantic revisionists attempt to resolve the Paradox in a more general manner by replacing (KP1) with some other principle which they suggest plays a theoretical role similar to that of (KP), but which can be shown to be consistent with either (NonOm1) or an appropriate reformulation thereof.

Relative to this taxonomy, the resolution I have pursued here is a form of semantic revisionism. It should be added, however, that since the term “re-

²²Brogaard and Salerno actually employ the term “semantic restrictionism” to describe the proposals to which I wish to compare the one described here. However, since these proposals are put forth by theorists who advocate reformulating (KP) based on considerations about in the intended interpretation of such a principle, the term revisionism seems equally apt.

visionism” is assessed relative to the formalization of (KP) and (NomOm) as (KP1) and (NonOm1), accepting this categorization does not mean we must acknowledge that neo-verificationism only can be salvaged only by revising (KP1). Rather, my strategy has been to examine the views of theorists who are the presumptive targets of the Knowability Paradox and to argue that (KP1) represents, at best, a simplistic means of formalizing their views. From this perspective it is thus Hart’s original use of (KP1) to formulate verificationism which constitutes an instance of revision.

I have already commented on the logical revisionist strategy in the course of addressing Williamson’s proposal that verificationism can be salvaged by arguing that intuitionistic logic should be employed when reasoning from (KP1) and (NonOm1).²³ As I have portrayed the neo-verificationists as staunch supporters of intuitionistic logic, there are reasons to be sympathetic to his desire to ground a resolution to the Paradox in considerations which have a basis in a detailed formulation of their views. Nonetheless, we have already seen that there are several reasons to think that the mere adoption of an intuitionistic background logic may be insufficient to adequately

²³A related proposal is that of Beall [4] and Wansing [110] who suggest that we use paraconsistent logic to reason about (KP1) and (NonOm1). Beall motivates his proposal on the basis of his claim the Knower Paradox provides independent motivation for accepting that the class of known statements is inconsistent. However, we have seen in Chapter 2 that there appears to be a resolution to the Knower which is both less radical and better motivated with respect to its intellectual origins.

resolve the Paradox. Thus since we also seen in the previous section that the Paradox can be resolved even if we reason classically, the approach adopted here provides no motivation for changing the background logic. And since this resolution also does not require restricting substitution into (KP), we feel the same way about the restrictionist strategy.²⁴

The issue to which I now wish to turn is that of comparing my proposed reformulation of (KP) to those of other semantic revisionists. The best known proposal in this class is due to Edgington [24].²⁵ Her proposal is grounded in the fact that we may distinguish between knowing that φ is true in a world w and knowing that φ would be true in another (potentially non-actual) world. On the basis of this observation, she proposes that (KP) be understood thusly:

(4.11) For all statements φ and worlds w_1 , if φ is true in w_1 , then there is a world w_2 in which it is known that φ is true in w_1 .²⁶

²⁴Dummett's [23] original resolution to the Paradox fits into this class. He proposed to limit substitution into (KP) to logically unstructured (or "basic") sentences. This obviously blocks the derivation of the Paradox since no instance of (NonOm) satisfies this criterion. However, the cost of this step is high, as it significantly limits the class of statements to which an epistemic constraint on truth can be asserted to apply.

²⁵Another proposal in this class is that of Kvanvig [63], [64]. His resolution is based on the observation that the occurrence of the possibility operator in (KP) creates a context in which the substitution of what he refers to as "modally non-rigid statements" (i.e. ones which may take on different truth values in different worlds) is not legitimate. Kvanvig's diagnosis of the locus of the Paradox is thus different from Edgington's. However, the revised version of (KP) which he formulates has problematic consequences similar to those incurred by Edgington's proposal which I will discuss in a moment.

In order to reason about the consequences of this principle, Edgington proposes that we employ an actuality operator A with the intended interpretation $A\varphi$ if and only if φ is true in the actual world. Her proposed reformulation of (KP) thus takes the form

(KP_e) For all φ , $A\varphi \rightarrow \Diamond KA\varphi$.

Suppose we additionally assume that there is a statement which is both true and unknown in the actual world – i.e.

(NomOm_e) For some ψ , $A(\psi \wedge \neg K\psi)$.

From this it follows that

$$(4.12) \quad \Diamond KA(\psi \wedge \neg K\psi)$$

which can be taken to express that there is some world w where $KA(\psi \wedge \neg K\psi)$ holds.

In order to assess the consequences of (4.12), we must additionally note that Edgington's intended interpretation of $K\varphi$ is “ φ is known by some agent i at some time t .” If we now let α denote the actual world, it follows

²⁶Edgington elects to explicate modal notions in terms of situations rather than worlds. She takes a situation to differ from a possible world in that the latter but not the former must be descriptively complete in settling the truth values of all propositions. Williamson [114], however, has argued that if the modal notions involved in the traditional exposition of the Paradox are understood in this way, then (KP_e) becomes trivial. I have thus modified Edgington's interpretation of A and $\Diamond$ to take this into account.

from (4.12) by a variant of the Fitch derivation that $w \neq \alpha$. Thus from (4.12) it follows that there must be non-actual agents who at some non-actual time have knowledge not just about the actual world α , but also about the epistemic situation of the agents who reside in it. Williamson has argued that since this consequence of (4.12) rivals the implausibility of (Om) itself, we must reject (KP_e) as a viable means of expressing an epistemic constraint on truth.

In order to appreciate Williamson's concern, consider what would be required in order for an agent i in a world $w \neq \alpha$ to entertain a thought which is uniquely about α . i cannot refer to α by using an indexical device akin to Edgington's A operator, as when uttered in w , statements of the form $A\varphi$ express that φ is true not in the actual world α , but rather in his own world w . And in virtue of common assumptions about the nature of modality, i is also barred from referring to α either by ostension or through any form of causal interaction. Thus despite whatever intuitions might be taken to motivate (KP_e) , this principle appears to lead to the conclusion that there must be "non-actual knowledge about the actual world."

As this consequence appears to be unacceptable, we must also ask whether the problem Williamson raises applies to the current proposal for resolving the Paradox. On the face of things, it might not seem like there is cause for

concern as there are no overtly modal or indexical devices in QLPE which could result in the sort of situation just described. However, recall the QLPE statement 3') which arises in parallel to (4.12) by substituting an instance of (NonOm3) for φ in (KP4):

$$3') \quad (\exists y)y : (\psi \wedge (\forall x)[E(x) \rightarrow \neg x : \psi])$$

As was discussed above, this statement reports that there is a proof which simultaneously serves to verify ψ and also the statement $(\forall x)[E(x) \rightarrow \neg x : \psi]$ which I have taken to formalize the proposition that ψ is not known. As there is no guarantee that such a proof will be the denotation of a closed term, let us temporarily introduce the term p to denote a verification with this property. It would then follow from 3') that $a_2 \cdot p : (\forall x)[E(x) \rightarrow \neg x : \psi]$. This formula asserts that $a_2 \cdot p$ is a proof of $(\forall x)[E(x) \rightarrow \neg x : \psi]$. And relative to the analysis of knowledge given in (4.2a), it then follows that if an agent were to have constructed $a_2 \cdot p$, he would thereby know $(\forall x)[E(x) \rightarrow \neg x : \psi]$. We must now ask whether this situation leads to any implausible consequences analogous to those entailed by (4.12).

Note first that Proposition 4.3.2 implies that $\neg E(a_1 \cdot p)$ – i.e. that the proof denoted by $a_1 \cdot p$ has not been constructed. It is not possible to derive an analogous conclusion about $a_2 \cdot p$ unless we also assume that $E(x)$ satisfies

some additional closure properties. But in this case, the conditions needed to derive $\neg E(a_2 \cdot p)$ are again quite plausible. For if we i) assume EAx , then it follows that ii) $E(a_1)$, from which iii) $\neg E(p)$ follows, and then since iv) $E(a_2)$ also follows from EAx , it follows that $\neg E(a_2 \cdot p)$ if we also assume $EDown$. For purposes of argument, it thus seems reasonable to assume that the proof denoted by $a_2 \cdot p$ has not been constructed and, as such, is a “merely potential” proof. Note, however, that this proof itself serves to verify a statement which can be taken to express that no proof of ψ has actually been constructed. If we also assume that all proofs in a model of QLPE are potentially constructible (and hence potentially graspable by a deductive agent), the existence of $a_2 \cdot p$ might be taken to entail something akin to the conclusion that there must be non-actual agents who have knowledge of the actual world.

Despite the similarity in form between (4.12) and 3'), the latter does not appear to incur the same sort of problems relative to the interpretation I have argued that the neo-verificationists should assign to the language of QLPE. In order to see this, we must examine in slightly more detail the intended interpretation of the predicate $E(x)$. I have thus far taken the notion of constructedness which $E(x)$ is intended to express largely for granted on the basis of the use of this notion by the neo-verificationists in framing the

distinction between actual and potential proofs which we discussed in Section 2. However, if we wish to say more about this concept, it is difficult to resist adopting an analysis resembling

(4.13) proof t is constructed $\iff$ there exists some agent who has
constructed t at or before the current time

The possibility of adopting a temporal interpretation of $E(x)$ emerges most directly from Prawitz's remarks on the nature of potential proof, which he repeatedly described as existing in an "abstract" or "tenseless" sense (cf., e.g., [84]). Actual proofs are thus presumably understood by Prawitz in contradistinction to potential ones in virtue of the fact that they have already been concretely produced. And it thus follows that a constructed proof can be considered to be a concrete entity existing in time.

If we adopt this interpretation of $E(x)$, we must also ask if it is contradictory to assume that there can exist potential proofs which verify statements about our actual (i.e. current) epistemic situation. In order to see why the existence of such proofs does not have counterintuitive consequences, it will be useful to temporarily employ an idealized model of deductive activity which was first employed in intuitionistic mathematics. Relative to this model, mathematical activity is represented by the operation of a single idealized

mathematical agent [IM], who comes to prove (or otherwise “experience the truth of”) various statements at a series of discrete temporal stages which are assumed to be indexed by natural numbers $0, 1, \dots$. Since the fact that IM has or has not proven a given statement at a particular stage corresponds to a contingent fact about his operation, it might first appear that the picture envisioned by this model is extraneous to the subject matter of mathematics itself. However Brouwer [11] made explicit use of reference to the IM in order to produce weak counterexamples to principles in classical analysis. In particular, he assumed that it was allowable to refer to stages at which IM has *not* proven certain statements.²⁷

The first attempt to formalize Brouwer’s methods was given by Kreisel [57] who proposed that the operation of IM could be axiomatized and then adjoined to a background system of intuitionistic analysis resulting in what came to be known as the Theory of the Creative Subject. This theory appears to represent the closest point of contact between an intuitionistic formal system and the neo-verificationists’ distinction between actual and potential proofs.²⁸ The precise formulation of this theory need not detain us here as many of its details pertain to specific ways in which the operation of IM

²⁷For a reconstruction of one of Brouwer’s so-called Creative Subject arguments which exhibits this property see [105], p. 842-843.

is assumed to be idealized. But there are two relevant senses the Creative Subject picture is *not* be an idealization of our actual deductive activity: 1) it allows that we may be ignorant of a statement now but come to know it later by constructing a proof; 2) it is consistent with conceiving of the construction of such a proof as a process which has duration.

These both seem to be undeniable features of our actual mathematical practices. For on the one hand, we conventionally speak of our mathematical knowledge as growing as we construct new proofs. And on the other, what we ultimately come to identify as a proof in practice is often the result of considerable effort on the part of a community of mathematicians whose work may proceed over a period of years. Thus it does not seem so far-fetched to think that proofs appear either with explicit dates attached (e.g. their dates of publication) or at least that the times at which they became constructed can be determined on the basis of *descriptive* (as opposed to purely indexical) features pertaining to the conditions under which they were produced.

²⁸For a discussion of the interpretative issues which arise when we attempt to formulate the Theory of the Creative Subject axiomatically, see [17]. Williamson [116] has also discussed this theory in the context of the Knowability Paradox. His primary conclusion is that the Paradox cannot be developed in the language which is commonly used to formulate this theory. This appears to be correct, because although this language provides a means of referring to the individual stages of IM's constructions, it provides no means of distinguishing the *current* stages from the others. However, the reason which Williamson gives for this – i.e. that the language of theory is intended to quantify only over “abstract” or “mathematical” times while the concept of knowability relevant to the Knowability paradox pertains to “concrete” or “non-mathematical” time – overlooks the fact that Brower employed the IM explicitly as a means of introducing reflection upon the actual

These observations suggest that we may legitimately take the construction of proofs to be *events* which, under only mild idealization, can be assumed to have a definite beginning, middle and end. This in turn suggests that it is allowable to employ what Burgess [14] has called *chronometry* – i.e. our normal means of dating events using calendrical terms like years, months, days, hours, etc. – as a canonical means of referring to the times at which the construction of proofs is completed (and at which they hence become constructed).²⁹ And for this reason, it does not appear to be necessary to assume that the mechanism by which $E(x)$ succeeds in picking out just those proofs which have been constructed at the current time must be indexical. Rather, it could function in a descriptive manner which is analogous to the function served by the use of stages in the Theory of the Creative Subject. Admittedly, the means by which reference to our current state of knowledge is achieved in practice is considerably more complex than the use of natural numbers to index discrete times. It will, for instance, be informed by knowl-

state of IM's knowledge into the practice of intuitionistic mathematics itself.

²⁹This observation highlights the connection between the way the neo-verificationists are likely to understand (KP) and its temporal analog – i.e. if φ is true, then φ *will* be known to be true – which Burgess refers to as the *Discovery Principle* (DP). One problem which Burgess suggests stands in the way of finding a plausible formulation of this principle is that of the applicability of chronometry to arbitrary events – i.e. not just ones involving the construction of proofs, but potentially durationless empirical phenomena like the collision of elementary particles. The foregoing considerations suggest that the applicability of chronometry is relatively unproblematic in the former case.

edge of a great many contingent facts about the course which our actual mathematical practice has taken in the past. But the fact that we are in possession of such a mechanism seems to be no more contentious than the claim that we can come to know that the sentence

(4.14) Every map is four colorable and it was not known before 1976 that
every map is four colorable.

expresses a true proposition by surveying the recent literature in graphing theory and topology.

Bibliography

- [1] C. A. Anderson. The paradox of the knower. *The Journal of Philosophy*, 80(6):338–355, 1983.
- [2] S. N. Artemov. Operational modal logic. Technical Report MSI 95–29, Cornell University, 1995.
- [3] S. N. Artemov. Explicit provability and constructive semantics. *The Bulletin of Symbolic Logic*, 7(1):1–36, 2001.
- [4] J. C. Beall. Fitch’s proof, verificationism and the knower paradox. *Australasian Journal of Philosophy*, 78(2):241–247, 2000.
- [5] A. Bezboruah and J. C. Shepherdson. Gödel’s second incompleteness theorem for Q. *The Journal of Symbolic Logic*, 41(2):503–512, 1976.
- [6] K. Binmore and H. Shin. Algorithmic knowledge and game theory. In C. Bicchieri and M. L. D. Chiara, editors, *Knowledge, Belief, and Strategic Interaction*, chapter 9, pages 141–154. Cambridge University Press, 1993.
- [7] G. Boolos. *The Logic of Provability*. Cambridge University Press, 1993.
- [8] G. Boolos, J. Burgess, and R. Jeffrey. *Computability and Logic*. Cambridge University Press, fifth edition, 2007.
- [9] G. Boolos and G. Sambin. Provability: the emergence of a mathematical modality. *Studia Logica*, 50:1–23, 1991.
- [10] B. Brogaard and J. Salerno. Fitch’s paradox of knowability. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2009 edition, 2009.

- [11] L. Brouwer. Essentieel negative eigenschappen [essentially negative properties]. *Indagationes math.*, 10, 1948.
- [12] L. Brouwer. Consciousness, Philosophy and Mathematics. In *Philosophy of Mathematics: Selected Papers (2nd ed.)*, pages 90–96. Cambridge University Press, 1983.
- [13] J. P. Burgess. Which modal logic is the right one? In *Mathematics, Models and Modality*. Cambridge University Press, 2008.
- [14] J. P. Burgess. Can Truth Out? In J. Salerno, editor, *New Essays on the Knowability Paradox*. Cambridge University Press, Forthcoming.
- [15] C. Cozzo. What Can We Learn From the Paradox of Knowability? *Topoi*, 13(2):71–78, 1994.
- [16] C. Cross. The paradox of the knower without epistemic closure. *Mind, New Series*, 110(438):319–333, 2001.
- [17] M. Dummett. *Elements of Intuitionism*. Oxford: Clarendon Press, 1977.
- [18] M. Dummett. The Philosophical Basis of Intuitionistic Logic. In *Truth and Other Enigmas*, pages 215–247. Oxford University Press, 1978.
- [19] M. Dummett. Realism. *Synthese*, 52(1):55–132, 1982.
- [20] M. Dummett. Reply to Prawitz. In B. M. Taylor, editor, *Michael Dummett: Contribution to Philosophy*, pages 117–165. Martinus Nijhoff Publishers, 1987.
- [21] M. Dummett. *The Logical Basis of Metaphysics*. Duckworth, 1991.
- [22] M. Dummett. What is a theory of meaning? (II). In *The Seas of Language*, pages 67–137. Oxford University Press, 1993.
- [23] M. Dummett. Victor’s error. *Analysis*, 61:1–2, 2001.
- [24] D. Eglington. The paradox of knowaiblity. *Mind*, 94:557–568, 1985.

- [25] P. Égré. The knower paradox in the light of provability interpretations of modal logic. *Journal of Logic, Language, and Information*, 14(1):13–48, 2005.
- [26] S. Feferman. Arithmetization of metamathematics in a general setting. *Fundamenta Mathematicae*, 49:35–92, 1960.
- [27] F. Fitch. A logical analysis of some value concepts. *Journal of Symbolic Logic*, 28(2):135–142, 1963.
- [28] F. Fitch. A goedelized conception of the prediction paradox. *American Philosophical Quarterly*, pages 161–64, 1964.
- [29] M. Fitting. Quantified LP. Technical report, CUNY Ph.D. Program in Computer Science Technical Report TR-2004019, 2004.
- [30] M. Fitting. Semantics and Tableaus for LPS4. Technical report, CUNY Ph.D. Program in Computer Science Technical Report TR-2004016, 2004.
- [31] M. Fitting. The logic of proofs, semantically. *Annals of Pure and Applied Logic*, 132(1):1–25, 2005.
- [32] M. Fitting. A quantified logic of evidence. *Annals of Pure and Applied Logic*, 152(1-3):67–83, 2008.
- [33] M. Fitting and R. L. Mendelsohn. *First-Order Modal Logic*. Kluwer Academic Publishers, 1998.
- [34] E. Gettier. Is justified true belief knowledge? *Analysis*, 23:121–123, 1963.
- [35] K. Gödel. An interperatation of intuitionistic propositional calculus. In S. Feferman et al., editors, *Kurt Gödel. Collected Works*, volume I, pages 144–195. Oxford University Press, 1986.
- [36] K. Gödel. On formally undecidable propositions of *Principia Mathematica* and related systems I). In S. Feferman et al., editors, *Kurt Gödel. Collected Works*, volume I, pages 144–195. Oxford University Press, 1986.

- [37] K. Gödel. Lecture at Zilsel's. In S. Feferman et al., editors, *Kurt Gödel. Collected Works*, volume III, pages 62–113. Oxford University Press, 1995.
- [38] K. Gödel. The present situation in the foundations of mathematics. In S. Feferman et al., editors, *Kurt Gödel. Collected Works*, volume III, pages 45–53. Oxford University Press, 1995.
- [39] A. Goldman. Causal theory of knowing. *The Journal of Philosophy*, 64:357–372, 1967.
- [40] N. D. Goodman. A theory of constructions equivalent to arithmetic. In A. Kino, J. Myhill, and R. E. Vesley, editors, *Proceedings Summer Conf. on Intuitionism and Proof Theory, Buffalo, NY, USA, Aug 1968*, Studies in Logic and the Foundations of Mathematics, pages 101–120. North-Holland, 1970.
- [41] N. D. Goodman. Epistemic arithmetic is a conservative extension of intuitionistic arithmetic. *J. Symb. Log*, 49(1):192–203, 1984.
- [42] I. Hacking. What is strict implication? *J. Symb. Log*, 28(1):51–71, 1963.
- [43] V. Halbach, H. Leitgeb, and P. Welch. Possible worlds semantics for modal notions conceived as predicates. *Journal of Philosophical Logic*, 32:179–223, 2003.
- [44] V. Halbach and P. Welch. Necessity and necessary truths: A prolegomenon to the use of modal logic in the analysis of intensional notions. *Mind*, 18:71–100, 2009.
- [45] S. Haldén. A pragmatic approach to modal theory. *Acta Philosophica Fennica*, 16(53-64), 1963.
- [46] M. Hand. Knowablity and epistemic truth. *Australasian Journal of Philosophy*, 81(2):216–228, 2003.
- [47] M. Hand. Performance and paradox. In *New Essay on the Knowability Paradox*. Oxford University Press, 2008.

- [48] W. Hart. The epistemology of abstract objects. *Aristotelian Society Supplementary Volume*, 53:163–165, 1979.
- [49] W. Hart and C. McGinn. Knowledge and necessity. *Journal of Philosophical Logic*, 5(2):205–208, 1976.
- [50] A. Heyting. A propos d’un article de MM. barzin et errera. *Enseignement Mathématique*, 31:121–122, 1932.
- [51] A. Heyting. *Intuitionism: An Introduction*. North-Holland, Amsterdam, 1956.
- [52] A. Heyting. *Mathematische Grundlagenforschung, Intuitionismus, Beweistheorie*. Springer, 1974.
- [53] J. Hintikka. *Knowledge and Belief*. Cornell University Press, 1962.
- [54] G. E. Hughes and M. J. Cresswell. *A New Introduction to Modal Logic*. Routledge, 1968.
- [55] R. Koons. *Paradoxes of belief and strategic rationality*. Cambridge University Press, 1992.
- [56] G. Kreisel. Foundations of intuitionistic logic. In E. Nagel, P. Suppes, and A. Tarski, editors, *LMPS*, pages 198–210. Stanford University Press, 1962.
- [57] G. Kreisel. Informal rigour and completeness proofs. In I. Lakatos, editor, *Problems in Philosophy of Mathematics*, pages 138–171. North Holland, 1967.
- [58] S. Kripke. On two paradoxes of knowledge. Manuscript.
- [59] S. Kripke. Semantical considerations on modal logic. *Acta Philosophica Fennica*, 24:83–94, 1963.
- [60] W. Künne. Two principles concerning truth. In R. Auxier and L. Hahn, editors, *The Philosophy of Michael Dummett*, pages 315–344. Open Court Press, 2007.

- [61] R. Kuznets. On decidability of the logic of proofs with arbitrary constant specifications (abstract). *Bulletin of Symbolic Logic*, 11:111, 2005.
- [62] R. Kuznets. *Complexity Issues in Justification Logic*. PhD thesis, CUNY Graduate Center, 2008.
- [63] J. Kvanvig. The knowability paradox and the prospects for anti-realism. *Noûs*, 29(4):481–500, 1995.
- [64] J. Kvanvig. *The Knowability Paradox*. Oxford: Clarendon Press, Oxford, 2006.
- [65] H. Leitgeb. On formal and informal provability. In *Waves in Philosophy of Mathematics*. Palgrave Macmillan, 2009.
- [66] D. Lewis. Elusive knowledge. *Australasian Journal of Philosophy*, 74:549–567, 1996.
- [67] S. Maitzen. The knower paradox and epistemic closure. *Synthese*, 114(337-354), 1998.
- [68] P. Martin-Löf. A Path from Logic to Metaphysics. In *Atti del Congresso Nuovi problemi della logica e della filosofia della scienza, Viareggio, 8-13 gennaio, 1990, Vol.II*, pages 141–149. Kluwer Academic Publishers, 1991.
- [69] P. Martin-Löf. Verificationism Then and Now. In DePauli-Schimanovich, W., et al., editor, *The Foundational Debate*, pages 187–196. Kluwer Academic Publishers, 1995.
- [70] P. Martin-Löf. On the meanings of the logical constants and the justifications of the logical laws. *Nordic Journal of Philosophical Logic*, 1(1):11–60, 1996. Text of lectures originally given in 1983 and distributed in 1985.
- [71] E. Martino and G. Usberti. Temporal and Atemporal Truth in Intuitionistic Mathematics. *Topoi*, 13:83–92, 1994.
- [72] V. McGee. *Truth, Vagueness and Paradox*. Hackett Publishing Company, 1991.

- [73] J. C. C. McKinsey and A. Tarski. Some theorems about the sentential calculi of Lewis and Heyting. *Journal of Symbolic Logic*, 13:1–15, 1948.
- [74] A. Mkrtychev. Models for the logic of proofs. In *Proceeding of LFCS*, pages 266–275, 1997.
- [75] R. Montague. Syntactic treatment of modalities, with corollaries on reflexion principles and finite axiomatizability. In R. Thomason, editor, *Formal Philosophy*. Yale University Press, 1974.
- [76] R. Montague and D. Kaplan. A paradox regained. *Notre Dame Journal of Formal Logic*, 1:79–90, 1960.
- [77] J. Myhill. Remarks on the notion of proof. *The Journal of Philosophy*, 57(14):461–471, 1960.
- [78] R. Parikh. Existence and feasibility in arithmetic. *Journal of Symbolic Logic*, 36(3):494–508, 1971.
- [79] P. Percival. Fitch and intuitionistic knowability. *Analysis*, 50:182–187, 1990.
- [80] A. Plantinga. *Warrant: the Current Debate*. Oxford University Press, 1993.
- [81] F. Pogliolesi. Three different solutions to the knower paradox. *Annali del Dipartimento di Filosofia dell’Università di Firenze*, 13:147–163, 2008.
- [82] D. Prawitz. Meanings and proofs: On the conflict between classical and intuitionistic logic. *Theoria*, 43(1):2–40, 1977.
- [83] D. Prawitz. Intuitionistic Logic : A Philosophical Challenge. In G. H. von Right, editor, *Logic and Philosophy*, pages 1–10. Martinus Nijhoff Publishers, 1980.
- [84] D. Prawitz. Dummett on a Theory of Meaning and its Impact on Logic. In B. M. Taylor, editor, *Michael Dummett: Contribution to Philosophy*, pages 117–165. Martinus Nijhoff Publishers, 1987.

- [85] D. Prawitz. Problems for a Generalization of a Verificationist Theory of Meaning. *Topoi: An International Review of Philosophy*, 64:87–92, 1998.
- [86] D. Prawitz. Truth and Objectivity from a Verificationist Point of View. In G. Dales and G. Oliveri, editors, *Truth in Mathematics*, pages 41–51. Oxford University Press, 1998.
- [87] W. V. Quine. Three grades of modal involvement. In *Proceedings of the XIth International Congress of Philosophy, Volume 14*, pages 65–81, 1953.
- [88] W. V. O. Quine. *Methods of Logic*. Holt, Reinhart and Winston, 1940.
- [89] P. Raatikainen. Conceptions of Truth in Intuitionism. *History and Philosophy of Logic*, 25(2):131–145, 2004.
- [90] H. Rasiowa and R. Sikorski. Algebraic treatment of the notion of satisfiability. *Fundamenta Mathematicae*, 40:62–95, 1953.
- [91] W. Reinhardt. Epistemic theories and the interpretation of Gödel’s incompleteness theorem. *Journal of Philosophical Logic*, 15(4):427–474, 1986.
- [92] S.A. Kripke. Outline of a theory of truth. *Journal of Philosophy*, pages 690–716, 1975.
- [93] J. Salerno. Knowablity noir. In *New Essays on the Knowability Paradox*. Oxford University Press, 2008.
- [94] S. Shapiro. Epistemic and intuitionistic arithmetic. In S. Shapiro, editor, *Intensional Mathematics*, pages 11–45. North Holland, 1985.
- [95] R. Shaw. The paradox of the unexpected exmination. *Mind*, 67:382–84, 1958.
- [96] R. Shope. Gettier problems. In E. Craig, editor, *Routledge encyclopedia of philosophy*, pages 3120–3125. Routledge, 1998.
- [97] B. Skyrms. An immaculate conception of modality or how to confuse use and mention. *Journal of Philosophy*, 75(7):368–387, 1978.

- [98] C. Smorynski. *Self-Reference and Modal Logic*. Springer-Verlag, 1985.
- [99] R. M. Solovay. Provability interpretations of modal logic. *Israel J. Math*, 25, 1976.
- [100] G. Sundholm. Constructions, Proofs and the Meaning of Logical Constants. *Journal of Philosophical Logic*, 12:151–172, 1983.
- [101] W. Tait. Gödel’s interpretation of intuitionism. *Philosophia Mathematica*, 14:208–228, 2006.
- [102] A. Tarski. The concept of truth in formalized languages. In A. Tarski, editor, *Logic, Semantics, Metamathematics*, pages 152–278. Oxford University Press, 1956. Translated by J.H. Woodger.
- [103] R. Thomason. A note of the syntactic treatment of modality. *Synthese*, 44:371–395, 1980.
- [104] A. Troelstra and H. Schwichtenberg. *Basic Proof Theory*. Cambridge University Press, 2000.
- [105] A. Troelstra and D. van Dalen. *Constructivism in Mathematics*, volume I, II. North Holland, 1988.
- [106] T. Tymoczko. An unsolved puzzel about knowledge. *Philosophical Quarterly*, 34:437–458, 1984.
- [107] G. Usberti. *Significato e Conoscenza: Per una Critica del Neoverificazionismo*. Guerini Scientifica, 1995.
- [108] M. van Atten. The development of intuitionistic logic. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2009 edition, 2009.
- [109] G. von Wright. *An essay in modal logic*. North Holland, 1951.
- [110] H. Wansing. Diamonds are a philosopher’s best friends. *Journal of Philosophical Logic*, 31:591–612, 2002.
- [111] S. Weinstein. The Intended Interpretation of Intuitionistic Logic. *Journal of Philosophical Logic*, 12:261–270, 1983.

- [112] P. Weiss. The prediction paradox. *Mind*, 61:265–269, 1952.
- [113] T. Williamson. Intuitionism disproved? *Analysis*, 42:203–207, 1982.
- [114] T. Williamson. On the paradox of knowability. *Mind*, 96:256–261, 1987.
- [115] T. Williamson. Two incomplete anti-realist modal epistemic logics. *The Journal of Symbolic Logic*, 55(1):297–314, Mar. 1990.
- [116] T. Williamson. On intuitionistic modal epistemic logic. *Journal of Philosophical Logic*, 21:63–89, 1992.
- [117] T. Williamson. *Knowledge and Its Limits*. Oxford University Press, 2000.
- [118] R. E. Yavorsky. Provability logics with quantifiers on proofs. *Annals of Pure and Applied Logic*, 113(1-3):373–387, 2001.