Solving the Knower Paradox via the Logic of Provability

Author: Mirjam A. de Vos
Student number: S1779974
Scientific discipline: Applied Mathematics

First supervisor: Prof. Dr. B.P. Kooi
Second supervisor: Prof. Dr. R. Verbrugge
Third supervisor: Dr. M. Pauly

Approved: February, 2015
Contents

1 Introduction 3
  1.1 What is a Paradox? ........................................ 3
  1.2 How to Solve a Paradox? .................................. 5
    1.2.1 Haack’s Requirements on a Solution to a Paradox . 6
    1.2.2 Tarski’s Hierarchy of Languages .................. 9
  1.3 Epistemic Logic ........................................... 11
  1.4 Different Formulations of the Knower Paradox .......... 13
    1.4.1 The Original Knower Paradox ....................... 13
    1.4.2 Cross’ Knower Paradox ............................... 15
    1.4.3 Cross’ Knowledge-Plus Knower Paradox ............ 16
  1.5 The Current Debate on the Knower Paradox ............. 17

2 Provability Logic 19
  2.1 Formal Arithmetic ....................................... 20
    2.1.1 The Language of Basic Arithmetic ................ 20
    2.1.2 Robinson Arithmetic ............................... 21
    2.1.3 Peano Arithmetic ................................... 22
    2.1.4 Peano Arithmetic and Modal Logic ................. 23
  2.2 Self-Reference in Arithmetic ............................ 24
    2.2.1 Gödel Numbering .................................. 25
    2.2.2 Diagonalization ................................... 27
  2.3 Gödel’s Incompleteness Theorems ....................... 28
  2.4 Löb’s Theorem and Derivability Conditions ............. 30
    2.4.1 Bernays-Hilbert-Löb Derivability Conditions .... 30
    2.4.2 Henkin’s Problem .................................. 31
    2.4.3 Löb’s Solution to Henkin’s Problem ............... 33

3 Egré’s Solutions to the Knower Paradox 36
  3.1 Skyrms’ Interpretation of Provability Logic ........... 37
  3.2 Anderson’s Interpretation of Provability Logic ......... 41
  3.3 Solovay’s Interpretation of Provability Logic ........ 43
  3.4 Summary .................................................. 47

4 The Quality of Egré’s Solutions 49
  4.1 The Quality of Skyrms’ Theory as a Solution ........... 49
    4.1.1 The Formal Part of Skyrms’ Theory as a Solution .. 49
    4.1.2 The Philosophical Part of Skyrms’ Theory as a Solution 50
    4.1.3 The Scope of Skyrms’ Theory as a Solution ....... 50
  4.2 The Quality of Anderson’s Solution ..................... 52
    4.2.1 The Formal Part of Anderson’s Solution .......... 52
    4.2.2 The Philosophical Part of Anderson’s Solution .... 53
    4.2.3 The Scope of Anderson’s Solution ................. 57
4.3 The Quality of Solovay’s Theory as a Solution
   4.3.1 The Formal Part of Solovay’s Theory as a Solution
   4.3.2 The Philosophical Part of Solovay’s Theory as a Solution
   4.3.3 The Scope of Solovay’s Theory as a Solution
4.4 Summary

5 Discussion and Conclusion
   5.1 Improving Égré’s Solutions?
   5.2 Interpreting Knowledge as Provability
   5.3 Conclusion

References
1 Introduction

This master’s thesis is about the knower paradox and certain solutions to it. Paul Egré [2005, p. 13] discusses “the kind of solution that modal provability logic provides to the [p]aradox” by surveying and comparing three different provability interpretations of modality. In this thesis, a lot of background is explained to clarify Egré’s solutions, where in particular attention is paid to the logic of provability or provability logic. To check whether Egré’s solutions are satisfactory solutions, we use the criteria of solutions to paradoxes defined by Susan Haack [1978]. This thesis aims to describe to what extent the knower paradox can be solved using provability logic.

Before we formulate the knower paradox, we look at the definition of a paradox in Section 1.1. In Section 1.2, we consider Haack’s requirements on a good solution, and we check whether Tarski’s hierarchy of languages is a satisfactory solution to self-referential paradoxes. Since the knower paradox is an epistemic paradox, we consider some basic epistemic logic in Section 1.3. Then we formulate the knower paradox in a few different ways in Section 1.4, and in Section 1.5 we consider the current philosophical debate on this paradox.

1.1 What is a Paradox?

The word paradox comes from the greek παραδοξός and δόξα, which mean something like ‘contrary to’ and ‘common opinion’. In the Stanford Encyclopedia of Philosophy, a paradox is considered as “a statement claiming something which goes beyond (or even against) ‘common opinion’ (what is usually believed or held)” [Cantini, 2012]. Another definition of paradoxes is from the book Paradoxes by Mark Sainsbury [1987, p. 1]. He writes that a paradox is “an apparently unacceptable conclusion derived by apparently acceptable reasoning from apparently acceptable premises”. This is the definition which we hold on to throughout the thesis, where we sometimes just state that the paradox is a certain ‘apparently unacceptable conclusion’.

We consider three paradoxes. Each of the following self-referential statements leads to a paradox that satisfies Sainsbury’s definition.

Statement (P1) is false. \hspace{1cm} (P1)

We know that statement (P2) is false. \hspace{1cm} (P2)

If statement (P3) is true, then C. \hspace{1cm} (P3)

Each one of the first two statements (Px) is used to create the apparently unacceptable conclusion that ‘(Px) is true if and only if (Px) is false’ (x = 1, 2). We assume the principle of bivalence, which states that every statement is either true or false.

3
(P1) The Liar Paradox We claim that statement (P1) is true if and only if (P1) is false, for (P1) as defined on Page 3. Suppose that (P1) is false, then by the contents of (P1), it does not hold that ‘statement (P1) is false’. Since everything is either true or false, we conclude that statement (P1) is true. This shows that (P1) is true if (P1) is false.

Suppose now that (P1) is true, then it holds that ‘statement (P1) is false’. This shows that (P1) is true only if (P1) is false. Since we have shown that the statement is false if it is true and true if it is false, we can conclude that (P1) is true if and only if (P1) is false.

(P2) The Knower Paradox Statement (P2) states ‘we know that statement (P2) is false’. The knower paradox is the apparently unacceptable conclusion that statement (P2) is true if and only if it is false.

Suppose (P2) is true. We assume that everything that is known is true\(^1\). Since statement (P2) states that ‘we know that statement (P2) is false’, it follows that statement (P2) is false. So if we suppose that the statement is true, then it follows that the statement is false. This is a contradiction, thus the assumption that statement (P2) is true cannot be true. Because this is the case, we infer that statement (P2) is false. Since we are the ones who inferred it, it follows that we know that (P2) is false\(^2\). However ‘we know that statement (P2) is false’ is exactly what the statement states, so the statement is true. So first it was shown that the statement is false if it is true, and then we inferred that it was false which implied that it was true. This means that (P2) is true if and only if (P2) is false.

(P3) Curry’s Paradox We have seen Sainsbury’s definition that a paradox is an apparently unacceptable conclusion derived by apparently acceptable reasoning from apparently acceptable premises. Curry’s paradox\(^3\) is the apparently unacceptable conclusion that any statement $C$ is true. For example\(^4\) the statement ‘Paris is the capital of Italy’ or ‘Santa Claus exists’ are statements that appear to be true if statement (P3), as defined on Page 3, is applied.

Why can we conclude $C$ using statement (P3)? Suppose (P3) is true. Then both (P3) itself and its antecedent are true, so the consequent $C$ is true. From this reasoning we conclude that ‘if statement (P3) is true, then $C$’. Since this is exactly what (P3) states, we conclude that (P3) is true.

---

\(^1\)This is a common assumption in epistemology, which we argue in Section 1.3.

\(^2\)In Section 1.3, we state the derivation rule ‘necessitation’, which allows us to derive ‘we know that $\phi$’ from ‘$\phi$ is true’.

\(^3\)In addition to Curry [1942], Löb [1955] mentioned this paradox. We call it Curry’s paradox, because he was the first and Löb mentioned it only as a remark by his referee. We meet Löb again in the next chapter.

\(^4\)These examples are extracted from respectively [Clark, 2002, p. 46] and [Boolos, 1995, p. 55].
And since ‘statement (P3) is true’, as the antecedent states, we infer C. This reasoning can be applied independently of the statement that is substituted for C, so we arrive at the apparently unacceptable conclusion that any statement C is true.

We have seen a few examples of paradoxes. The paradoxes that followed from (P1) and (P3) are semantic paradoxes, in which both mathematical and linguistic elements are involved (group A of ‘contradictions’ by Ramsey [1925, p. 20]). The knower paradox is an epistemic paradox, which is relevant for epistemology [Sorensen, 2014]. In this thesis, we do not consider other kinds of paradoxes, for example set-theoretic paradoxes in which no linguistic part is involved (group B by Ramsey).

In the current section, we saw an informal version of the knower paradox. Henceforth, we consider a paradox to be defined in a formal theory or system, consisting of a set of theorems. In the next section, we explain what it means to solve a paradox and some requirements on a good solution are considered.

1.2 How to Solve a Paradox?

We have seen what paradoxes are and three examples were considered. As a reminder, a statement is paradoxical if it is an apparently unacceptable conclusion derived by apparently acceptable reasoning from apparently acceptable premises.

There are two different kinds of solutions to paradoxes. A paradox is solved if we discard one of the axioms or rules of inference and accept the resulting theory in which the ‘apparently unacceptable conclusion’ cannot be derived. A paradox is also solved if in the new theory the conclusion can again be formulated but is not ‘apparently unacceptable’, like it was in the old system, then the paradox is also solved. An example of a theory which solves certain paradoxes in this second way is dialetheism, which is the view that there are true statements of the form ‘(Px) is true if and only if (Px) is false’ [Priest and Berto, 2013]. Many conclusions which are in other systems considered as paradoxes are not paradoxical in this system. In this thesis, we focus on the first kind of solutions, in which the ‘apparently unacceptable conclusion’ cannot be derived.

In this section, we discuss some requirements that make a solution to a paradox satisfactory (Section 1.2.1), after which we check whether a certain solution which abandons self-reference is a good solution for certain paradoxes (Section 1.2.2). After that a short review of epistemic logic is presented (Section 1.3). Then we state some different formulations of the knower paradox (Section 1.4) and consider its current status (Section 1.5).

\[5\] In this thesis, we consider a system and a theory as coinciding concepts.
1.2.1 Haack’s Requirements on a Solution to a Paradox

In her *Philosophy of Logics*, Susan Haack [1978] describes three requirements on solutions to paradoxes. First, a solution should give a consistent formal theory. This theory should indicate which of the premises or principles of inference from the theory in which the paradox is formulated should be disallowed. The second requirement is that a solution should give a philosophical explanation of why this premise or principle of inference seems acceptable but is unacceptable. Finally, there is the requirement that a solution should not be too broad or too narrow. We consider these requirements in more detail.

(1) The Formal Part of a Solution  
According to Haack [1978, p. 138–139], a solution to a paradox “should give a consistent formal theory (of semantics or set theory as the case may be) - in other words, indicate which apparently unexceptionable premises or principle of inference must be disallowed (the *formal* solution)”. So suppose we want to solve the Liar paradox, then we need a consistent formal theory \( \Sigma \) which does not contain the paradox. Since the paradox exists in the formal theory in which it is formulated, there is a difference between that theory and the consistent theory. This difference indicates which apparently acceptable premises or principles of inference are the ones that should be disallowed. Because in this thesis, we consider formal systems consisting of theorems based on axiom schemes, we add that a system which solves a paradox can also indicate a set of apparently acceptable theorems which should be disallowed. This system is consistent if \( \Sigma \vdash \bot \) does not hold.

The system in which the paradox is formulated consists of a set of theorems, defined by premises and rules of inference. By forming a new system, in which one of the premises or rules of inference from the old system is rejected, we arrive at a new set of theorems. Except the ‘apparently unacceptable conclusion’, there might be other theorems which are derivable in the old system, but not in the new one. We explain which requirements should be met by this new set of theorems when we describe Haack’s third requirement.

(2) The Philosophical Part of a Solution  
After stating that the formal solution to a paradox should contain a formal theory that indicates an apparently acceptable premise or principle of inference that should be disallowed, Haack [1978, p. 139] continues that a solution should “supply some explanation of why that premise or principle is, despite appearances, exceptionable (the *philosophical* solution)”’. This explanation should show that “the rejected premise or principle is of a kind to which there are (...) objections independent of its leading to paradox”.

Suppose we have a formal theory in which the liar paradox exists, and
we replace this by a new theory which only differs from the original one by disallowing the statements that mean the same as “this statement is false”. The only reason why we say these statements should be disallowed is ‘because they result in a paradox’. This is a solution that does not satisfy Haack’s philosophical criterion. We need to find philosophical arguments for disallowing apparently acceptable principles of inference and premises in order to have a satisfactory solution.

(3) The Scope of a Solution  A solution to a paradox is required to have the right scope, which means that it should not be too broad or too narrow. A solution is too broad if it is “so broad as to cripple reasoning we want to keep” [Haack, 1978, p. 139], and it is too narrow if it does not block all paradoxes which are closely related to the paradox which is solved. It is disputable what exactly those closely related paradoxes are. For example, if the solution solves a paradox of the form ‘P if and only if ¬P’, then other paradoxes of this form are considered as closely related to the solved paradox.

We can explain this in a more formal way. Suppose we consider a certain solution to a given paradox. Remember that there are two sets of theorems, namely the one from the system $S_1$ in which the paradox is present and the one from the system $S_2$ which is proposed as solution to the paradox. Consider the set $S$ as the union of these two sets of theorems, which are derived from ‘apparently acceptable premises and principles of inference, because $S_1$ and $S_2$ are based on those. We divide set $S$ into two subsets $A$ and $B$, where $A$ and $B$ are independent of $S_1$ and $S_2$. Set $B$ contains the paradox itself, together with all other ‘apparently unacceptable conclusions’ that occur in $S_1$ or in $S_2$. All other theorems of $S$ are in $A$. A solution of a good scope would reject exactly all statements from set $B$ (Figure 1d).

The solution is too broad if it rejects a statement from set $A$ (see Figure 1a) and it is too narrow if it does not reject all of the statements of set $B$ (see Figure 1b). A solution can be both too broad and too narrow, namely if a statement from set $A$ is rejected while at the same time some statement from set $B$ is not (see Figure 1c).

Suppose some system which is proposed to solve the liar paradox does not contain this paradox, but for some reason does contain a paradox based on the following two consecutive sentences: ‘The next sentence is true. The previous sentence is false’\(^6\). Both paradoxes are in set $B$, but the second one is not rejected in the new system. This means that the solution is too narrow. If for example another solution implies that a sentence like ‘this sentence is true’, which is in set $A$, cannot be true, then this solution is too broad.

\(^6\)The paradox is the apparently unacceptable conclusion of the form ‘$P$ is true if and only if $P$ is false’, where statement $P$ is one of the two sentences.
Figure 1: Possible scopes of a solution. The wave areas indicate the set of theorems which are rejected by the solution.

How is this requirement on the scope of a solution related to the requirements on its formal and its philosophical part? According to Haack [1978, p. 139], the requirement that a solution should not be too narrow "urges simply that the solution be such as to restore consistency" at the formal level. So, if we conclude that some solution to some paradox is consistent, which is necessary to satisfy Haack’s first requirement, then this solution is automatically not too narrow. Consider some formal system in which each ‘apparently unacceptable conclusion’ that occurs in the system is represented by ‘⊥’ or implies that ‘⊥’ is a theorem. It is obvious that a consistent solution to this paradox does not allow any theorem from the set $B$ of ‘apparently unacceptable conclusions’. So such a solution is not too narrow.

If an inconsistent system is presented as formal part of a solution to some paradox, then this solution is too narrow. This is the case, because it contains the paradox consisting of the ‘apparently unacceptable conclusion’ that the solution, which contains only ‘apparently acceptable premises and principles of inference’, is inconsistent.

Whether the formal part of the solution is satisfactory does not depend on whether the solution is too broad or not. If a solution is too broad, it
disallows statements which are not paradoxical, but it can still be consistent. If it is not too broad, it can be consistent, but it can also happen that it is inconsistent.

We consider the requirements on the scope of a solution with respect to its philosophical part. The requirement that a solution should not be too narrow “urges that the explanation offered [goes] as deep as possible” at the philosophical level [Haack, 1978, p. 139]. Remember that a solution is too narrow if it does not block all paradoxes which are closely related to the paradox which is solved. The deeper the explanation for disallowing certain premises or principles of inference goes, the more paradoxes which are not ‘closely related’ to the solved paradox are solved to. The more paradoxes disappear by the explanation, the better the philosophical part of the solution is, but of course a system does not necessarily include definitions of all kinds of concepts. For example a system in which sets are not defined is not viewed as solving any set-theoretic paradoxes. A system that solves different kinds of paradoxes is preferred, but is not necessary. We say a solution is not too narrow, even if it does not solve all paradoxes, but only those which are in a weaker sense closely related to the solved paradox.

If a solution is too broad, this indicates that there is something wrong with the philosophical part of the solution. The philosophical part of the solution contains arguments for rejecting a certain premise or principle of inference. This means that it should contain reasons to disallow all statements which are rejected by the formal part of the solution. If the solution appears to be too broad, apparently, some statements which are actually acceptable are rejected. This implies that the arguments for disallowing the premise or principle of inference are not sufficient. So if a solution is too broad, it does not contain a sufficient philosophical part. The other way around does not hold. A solution can be not too broad, while the philosophical part is insufficient or even lacking.

Summarizing, a solution to a paradox satisfies Haack’s criteria if it has an appropriate formal part and a satisfactory philosophical part and if it is not too broad or too narrow. We now consider one possible solution for every self-referential paradox, namely Tarski’s hierarchy of languages. We discuss to what extent this solution satisfies Haack’s criteria in the same way as three interpretations of provability logic are evaluated as solutions to the knower paradox in Chapter 4.

1.2.2 Tarski’s Hierarchy of Languages

A solution to any self-referential paradox might be to abandon self-reference. We consider an example of such a solution, namely the hierarchy of languages by Alfred Tarski [1930/1956]. We find out whether this solution satisfies
Haack’s criteria after we consider the idea of the hierarchy of languages.

In Tarski’s hierarchy of languages there is one language $L_0$, which is called the object language. We cannot say ‘$p$ is true in $L_0$’ for $p$ a sentence in language $L_0$, but there is a meta-language $L_1$ which has a truth predicate that applies to $L_0$. To say whether a sentence in language $L_1$ is true, we need another meta-language $L_2$ with a new truth predicate. In this way, every language $L_i$ gets a meta-language $L_{i+1}$ with a truth predicate that applies only to the sentences of $L_j$ for $j \leq i$ [Bolander, 2014, Section 3.1].

This hierarchy of languages does not allow that a sentence in language $L_i$ says something about its own truth, because it can only say something about the truth of sentences in languages $L_0$ up to and including $L_{i-1}$. Consider again statement (P2), which is stated in Section 1.1. Statement (P2) states that ‘we know that statement (P2) is false’ and leads to the knower paradox. Applying the hierarchy of languages to (P2), we see the following. If (P2) is a sentence in language $L_i$, then the truth of (P2) can only be expressed in a meta-language $L_j$ for $j > i$. This means that the content of (P2), namely the part which says that statement (P2) is false, cannot be expressed in $L_i$. Because we supposed that (P2) was a sentence in language $L_i$ and we inferred that this cannot be the case, we conclude that the sentence cannot be formulated in Tarski’s hierarchy of languages.

To what extent does this solution to self-referential paradoxes like the knower paradox satisfy Haack’s criteria that were discussed in Section 1.2.1? The formal part of the solution is the hierarchy of languages that indicates that the idea that there is only one language level should be disallowed. Is Tarski’s hierarchy consistent? Consider the following pair of sentences.

(A) Sentence B is true.
(B) Sentence A is false.

Suppose sentence A is a sentence in language $L_i$. Since A says something about the truth of sentence B, B should be in language $L_{i-1}$ or a lower language. Since B says something about the truth of A, it should be the case that sentence A is at language $L_{i-2}$ of lower. This means sentence A is in language $L_i$ as well as in language $L_{i-2}$ or lower, which is not allowed because a sentence is defined within a certain language and a unique definition of sentence A is required. Individually, sentence A can be formulated in Tarski’s hierarchy of languages, because there is no problem with the combination of statements A and B if statement B for example states ‘Rome is the capital of Italy’. The same holds for sentence B as described above, because we can define B as ‘sentence A is false’ in a unique language if statement A states for example ‘Rome is the capital of Italy’. There arises a problem if A and B are stated together in the way that is done above, because then the statements do not have a unique language level anymore.
This means that there is something wrong with the formal theory, so Haack’s first requirement is not satisfied.

Does the solution contain a satisfactory philosophical part, or is the reason to accept the hierarchy just the fact that this new theory solves the paradox? We did not mention it, but Tarski came up with a theorem on the undefinability of truth which might be sufficient as philosophical argument for accepting his hierarchy of languages. We do not discuss this theory here, but we argue why Haack’s third criterion is not satisfied. The sentence ‘This sentence is true’, cannot be formed in Tarski’s hierarchy of languages for the same reason why (P2) can’t. This means Tarski’s solution is too broad.

So abandoning self-reference in the way that is done by Tarski is not a satisfactory solution, because it is too broad and because there are sentences that can be formulated individually but not together.

We have discussed what paradoxes are and considered some requirements a solution to a paradox should satisfy. In the next section, we consider some theory of epistemic logic. Then we look at three different formulations of the knower paradox, after which we shortly consider the discussion about this paradox that is going on nowadays.

1.3 Epistemic Logic

Since the knower paradox is an epistemological paradox, we discuss a few ideas from epistemic logic. Epistemic logic focuses on propositional knowledge and belief. We do not consider propositional belief in this thesis. Propositional knowledge is knowledge which is about propositions, so we consider sentences like “I know that the battery of my laptop has run down” and not sentences like “I know you” or “I know that language”.

Although An Essay in Modal Logic by Georg H. Von Wright [1951] was important for the development of modal and epistemic logic, the book Knowledge and Belief: An Introduction to the Logic of the Two Notions by Jaakko Hintikka [1962] is “the first book-length work to suggest using modalities to capture the semantics of knowledge” [Rybakov, 2014, p. 2]. An important part of epistemic logic are the so-called Kripke models \( \langle S, \pi, R_1, \ldots, R_m \rangle \) for \( m \) agents, named after their inventor Saul A. Kripke [1959]. We assume that theory about these Kripke models is known, but we state the definition of the knowledge predicate and briefly consider two modal systems that are widely used to represent knowledge. The statement ‘agent \( i \) knows \( \phi \) in world \( \langle M, s \rangle \) if and only if \( \phi \) is true in all worlds that \( i \) considers possible’ is denoted by \( \langle M, s \rangle \models K_i \phi \), where the operator \( K_i \) uses sentences \( \phi \) as input. The statement is formally expressed as follows [Meyer and van der Hoek, 2004].

\[
\langle M, s \rangle \models K_i \phi \text{ if and only if } \langle M, t \rangle \models \phi \text{ for all } t \text{ with } (s, t) \in R_i
\]
So for some model $M$, $K_i \phi$ is true in world $(M, s)$ if $\phi$ is true in all possible worlds that are accessible from $(M, s)$ via $R_i$. Axiom schemes that are used in the knowledge systems we consider are the following [Meyer and van der Hoek, 2004, p. 13, 23].

1. All (instances of) propositional tautologies
   (A1)
   
   \[
   (K_i \phi \land K_i (\phi \to \psi)) \to K_i \psi
   \]
   for $i = 1, \ldots, m$

2. $K_i \phi \to \phi$
   (A2)
   for $i = 1, \ldots, m$

3. $K_i \phi \to K_i K_i \phi$
   (A3)
   for $i = 1, \ldots, m$

4. $\neg K_i \phi \to K_i \neg K_i \phi$
   (A4)
   for $i = 1, \ldots, m$

We briefly explain what the axiom schemes say. Scheme (A2) states that if agent $i$ knows $\phi$ and she knows $\phi$ implies $\psi$, then agent $i$ knows $\psi$. Axiom schemes (A4) and (A5) represent, respectively, positive and negative introspection: An agent knows that she knows something and she knows that she does not know something. Finally, axiom scheme (A3) states that knowledge implies truth, which is used in Section 1.1 to explain why statement (P2) (‘We know that statement (P2) is false.’) implies the knower paradox. It is a common assumption in epistemology that knowledge implies truth, which we argue via three quotations. At first, Hintikka [1962, p. 43] thinks it is “obvious that [this] condition has to be imposed on model sets”. The Stanford Encyclopedia of Philosophy states the following as part of the definition of knowledge of propositions as justified true belief. “False propositions cannot be known. Therefore, knowledge requires truth” [Steup, 2014, Section 1.1]. The same principle is stated by Lenzen [1980, p. 52]: “gewußt werden kann nur, was auch wahr ist”.

We consider two axiom systems in which at least some of the mentioned axiom schemes are used. Axiom system $K$ consists of (A1) and (A2), and of the following derivation rules.

\[
\frac{\phi, \phi \to \psi}{\psi}
\]

(R1)

\[
\frac{\phi}{K_i \phi}
\]

for $i = 1, \ldots, m$

(R2)

The first rule is called modus ponens and the second one necessitation.

A second knowledge system is $S_5$, which is an extension of $K$. It consists of (A1) up to and including (A5) and the derivation rules (R1) and (R2). In some articles, axiom scheme (A2) is called $K$, (A3) corresponds to $T$, (A4) to $4$, and (A5) to $5$ (for example in [Hendricks and Symons, 2014] and [Egré, 2005]).
We use these ideas about epistemic logic in the next section, where we formulate the knower paradox in three different ways before we close this chapter by exploring the current debate on this paradox.

1.4 Different Formulations of the Knower Paradox

We have already seen an informal version of the knower paradox, which was the conclusion ‘statement (P2) is true if and only if it is false’. Here (P2) is the statement ‘we know that statement (P2) is false’. We can also describe this paradox in a formal theory, which we do in three different ways. The first one is the original formulation of the knower paradox by Kaplan and Montague [1960], while the other two are formulated four decades later by Cross [2001]. We also mention two paradoxes with which this knower paradox should not be confused.

1.4.1 The Original Knower Paradox

The first formulation of the knower paradox is given by David Kaplan and Richard Montague. They use elementary syntax, by which they understand “a first-order theory containing (...) all standard names (of expressions), means for expressing syntactical relations between, and operations on, expressions, and appropriate axioms involving these notions” [Kaplan and Montague, 1960, Footnote 10, p. 89]. Besides that, it is given that ‘A ⊩ B’ means that B is derivable from A within the elementary syntax and ‘⊢ A’ means that A is provable within this syntax. In addition we give names to expressions, where we denote φ as the name of expression φ. These names can be defined via Gödel numbering, which we discuss in Section 2.2.1. Using this, it is possible to create self-referential arithmetical statements. The following two formulae are added to the elementary syntax.

\[ K(\bar{\phi}) \quad A \text{ knows the expression } \phi \text{ (on Sunday afternoon)}^{7} \]

\[ I(\bar{\phi}, \bar{\psi}) \quad \phi \vdash \psi \]

Remember that in Section 1.3, \( K_\phi \) was considered as an operator that used sentences \( \phi \) as input. We call this a sentential operator. A predicate \( K(\bar{\phi}) \) with sentence name \( \bar{\phi} \) as input is called a metalinguistic predicate. In both cases the result of an execution of an operation or a predicate is a sentence. We consider the following decree: “A knows that the present decree is false”. According to Kaplan and Montague [1960, p. 87], we can regard some sentence D as expressing this decree, namely D satisfying

\[ \vdash D \equiv K(\neg D). \]

\[ ^7 \] The ‘on Sunday afternoon’ part is there because the knower paradox arose from the hangman or unexpected examination paradox. We leave this part out from now on. Kaplan and Montague use \( K_S \), \( K \), and \( D \) where we use \( K \), \( A \), and \( D \) respectively.
From this expression some version of the knower paradox is derived, if the following three assumptions are made.

\[
E_1 := K(\neg D) \rightarrow \neg D \tag{E1}
\]

\[
E_2 := K(E_1) \tag{E2}
\]

\[
E_3 := [I(E_1, \neg D) \land K(E_1)] \rightarrow K(\neg D) \tag{E3}
\]

These premises are apparently acceptable. The assumption \(E_1\) says that if \(A\) knows the expression \(\neg D\), then \(\neg D\) is true. This corresponds to the idea that something false cannot be known. It is an instance of axiom schema (A3) from epistemic logic adjusted to names of sentences as input instead of sentences themselves. Assumption \(E_2\) expresses that assumption \(E_1\) is known by \(A\). It is a common premise that \(A\) knows that what she knows is true, and \(E_2\) just expresses that this is the case for knowing \(\neg D\). Finally, \(E_3\) expresses that if \(\neg D\) is derivable from \(E_1\) and \(A\) knows \(E_1\), then \(A\) knows \(\neg D\). This is an example of the epistemic closure principle, which we discuss in Section 1.5. It is not an instance of axiom schema (A2), because \(I(\phi, \psi)\) does not correspond to \(K_i(\phi \rightarrow \psi)\).

From these assumptions \(E_1, E_2\) and \(E_3\), the knower paradox can be derived as the following apparently unacceptable conclusion ‘\(\vdash D \leftrightarrow \neg D\)’. We derive this as follows, where we use the notation of derivation rules like hypothetical syllogism (HS) and PC for Propositional Calculus by Meyer and van der Hoek [2004]. We denote ‘\(E_i \vdash \phi\)’ in proof line (\(j\)) if the definition of \(E_i\) is used to derive the statement in line (\(j\)) or a statement in one of the former lines (1), (2), . . . , (\(j - 1\)).

\[
\begin{align*}
(1) \quad & \vdash D \leftrightarrow K(\neg D) \quad \text{by definition}^8 \text{of } D \\
(2) \quad & \vdash D \rightarrow K(\neg D) \\
(3) \quad & E_1 \vdash K(\neg D) \rightarrow \neg D \quad \text{by definition of } E_1 \\
(4) \quad & E_1 \vdash D \rightarrow \neg D \quad \text{by (2), (3), HS} \\
(5) \quad & E_1 \vdash \neg D \quad \text{by (4), PC} \\
(6) \quad & E_1 \vdash I(E_1, \neg D) \quad (5), \text{by definition of } I \\
(7) \quad & E_1, E_2 \vdash K(E_1) \quad \text{by definition of } E_2 \\
(8) \quad & E_1, E_2 \vdash I(E_1, \neg D) \land K(E_1) \quad \text{by (6), (7), PC} \\
(9) \quad & E_1, E_2, E_3 \vdash [I(E_1, \neg D) \land K(E_1)] \rightarrow K(\neg D) \quad \text{by definition of } E_3 \\
(10) \quad & E_1, E_2, E_3 \vdash K(\neg D) \quad \text{by (8), (9), R1} \\
(11) \quad & E_1, E_2, E_3 \vdash K(\neg D) \rightarrow D \quad \text{by (1), EE} \\
(12) \quad & E_1, E_2, E_3 \vdash D \quad \text{by (10), (11), R1} \\
(13) \quad & E_1, E_2, E_3 \vdash \neg D \rightarrow D \quad \text{by (5)-(12), PC} \\
(14) \quad & E_1, E_2, E_3 \vdash D \leftrightarrow \neg D \quad \text{by (4), (13), EI}
\end{align*}
\]
Another way of formulating an apparently unacceptable conclusion from the assumptions and the definition of $D$ is leaving out (13) and (14) and concluding $\vdash \bot$ from (5) and (12). In both ways, the paradox is used to prove that a system in which assumptions $E_1$, $E_2$, and $E_3$ are made is inconsistent. We consider two other formulations of the knower paradox, by which again it is concluded that a certain set of assumptions is inconsistent.

1.4.2 Cross’ Knower Paradox

Charles Cross [2001] presents a version of the knower paradox in which assumption $E_3$ from Kaplan and Montague [1960] is dropped. To do this, he defines a predicate $K'(\overline{\sigma})$ as follows, where $I(\overline{y}, \overline{x})$ is equivalent to '$y \vdash x$', like it was in the original formulation of the knower paradox in Section 1.4.1.

$$K'(\overline{x}) := \exists y(K(\overline{y}) \land I(\overline{y}, \overline{x}))$$

While $K(\overline{x})$ means 'A knows the expression $x$', $K'(\overline{x})$ means '$x$ is derivable from something known by A'. In Section 1.4.1, three single assumptions were made, but we use axiom schemes from now on. This means that we use a variable, namely $\phi$, which can be substituted by any statement from the language of the system. Cross makes the following two assumptions using the new predicate $K'$.

$$E_1' := K'(\overline{\sigma}) \rightarrow \phi \quad (E_1')$$
$$E_2' := K'(K'(\overline{\sigma}) \rightarrow \phi) \quad (E_2')$$

We expect an assumption $E_3'$ if we compare this set of assumptions with the one from the original formulation of the knower paradox. We do not need to assume this, but we can derive $E_3'$, which is the following statement.

$$E_3' := [I(\overline{\sigma}, \overline{\psi}) \land K'(\overline{\sigma})] \rightarrow K'(\overline{\psi}) \quad (E_3')$$

This can be derived via the following reasoning, where $I(x, y)$ is assumed to be transitive ($I(x, y) \land I(y, z) \rightarrow I(x, z)$ for all $x, y, z$).

\begin{align*}
(1) & \vdash K'(\overline{\sigma}) \rightarrow \exists y(K(\overline{y}) \land I(\overline{y}, \overline{\sigma})) & \text{by definition of } K' \\
(2) & \vdash K'(\overline{\sigma}) \land I(\overline{\sigma}, \overline{\psi}) \rightarrow \exists y(K(\overline{y}) \land I(\overline{y}, \overline{\sigma}) \land I(\overline{\sigma}, \overline{\psi})) & \text{by (1), PC} \\
(3) & \vdash K'(\overline{\sigma}) \land I(\overline{\sigma}, \overline{\psi}) \rightarrow \exists y(K(\overline{y}) \land I(\overline{y}, \overline{\psi})) & \text{by (2), transitivity of } I \\
(4) & \vdash K'(\overline{\sigma}) \land I(\overline{\sigma}, \overline{\psi}) \rightarrow K'(\overline{\psi}) & \text{by (3), definition of } K'
\end{align*}
This shows that \( E_3' \) follows from \( E_1' \) and \( E_2' \). The knower paradox can be derived from \( E_1', E_2', \) and \( E_3' \) in the same way as it can be derived from \( E_1, E_2, \) and \( E_3, \) if we substitute \( \phi \) by \( \neg D \). The difference is that in Cross’ theory, \( E_3' \) is not needed as an axiom. One other formulation of the knower paradox is left to discuss. This one is again from Cross [2001] and the definition of \( K' \) is applied there too.

### 1.4.3 Cross’ Knowledge-Plus Knower Paradox

The second formulation of the knower paradox which Cross [2001] published (in the same article as the first one) is called the paradox of the knowledge-plus knower. As definition of \( K'(\neg x) \), the one for the first knower paradox formulation by Cross (see Section 1.4.2) is used. The paradox of the knowledge-plus knower is named after the term knowledge-plus, which Cross uses to refer to sentences picked out by \( K' \), in the same way as sentences picked out by \( K \) are referred to as knowledge. The following premises are assumed.

\[
\begin{align*}
E_1' &:= K'(\overline{\phi}) \rightarrow \phi & (E_1') \\
E_2'' &:= K(K'(\overline{\phi}) \rightarrow \phi) & (E_2'')
\end{align*}
\]

The previous formulation of the knower paradox is used to prove that the current formulation leads to an unacceptable conclusion too. The only difference between the two sets of premises is that in \( E_2'' \), \( K \) is the main predicate, where in \( E_2 \), this is \( K' \). To derive the paradox we consider sentence \( D \) again, which satisfies \( \vdash D \iff K(\neg D) \). We use the following abbreviations.

\[
\begin{align*}
E_1'd &:= K'(\overline{\neg D}) \rightarrow \neg D & (E_1'd) \\
E_2'd &:= K'(\overline{E1'd}) & (E_2'd) \\
E_2''d &:= K(\overline{E1'd}) & (E_2''d)
\end{align*}
\]

We abbreviate to \( Xd \) for \( X = E_1', E_2', E_3' \), because \( Xd \) is the instance of \( X \) for \( \phi = \neg D \). Remember that Cross’ knower paradox (Section 1.4.2) concluded that \( E_1' \) and \( E_2' \) lead to an inconsistency. For \( \phi = \neg D \), it follows via the same reasoning that \( E_1'd \) and \( E_2''d \) lead to an inconsistency, which we denote by ‘\( E_1'd, E_2''d \vdash \bot \).’

In the defined syntax, it is the case that ‘\( \vdash I(\overline{\pi}, \overline{\pi}) \)’ is true. Notice that this implies that ‘\( E_2''d \vdash E_2'd \)’ by the following reasoning. It is the case that \( K(E_1'd) \) implies \( K(E_1'd) \land I(E_1'd, E_1'd) \). This means that there exists a \( y \) such that \( K(y) \) and \( I(y, E_1'd) \), namely \( y = K(E_1'd) \). This means that by definition of \( K' \), we have \( K'(E_1'd) \), so ‘\( K(E_1'd) \vdash K'(E_1'd) \)’, and thus
′E2′d ⊨ E2′d′, holds. The paradox can be derived now as follows.

(1) \(E1′d, E2′d ⊨ ⊥\) by Cross’ knower paradox
(2) \(E2′′d ⊨ E2′d\) as argued above
(3) \(E1′d, E2′′d ⊨ ⊥\) by (1), (2)

We have seen three formal formulations of the knower paradox now. They differ in the assumptions that are made, but all are formulated in the form of a set of premises about knowledge and result in the conclusion that the theory that supports that set of premises is inconsistent.

There are some paradoxes with which the knower paradox, as discussed in this thesis, should not be confused. One example is the knowledge or knowability paradox by Fitch [1963]. This knowledge paradox concludes that knowledge is equivalent with truth by assuming the verification thesis that everything that is true can be known. See [de Vos, 2013] to find out whether a certain solution to this paradox satisfies Haack’s criteria.

Another paradox concludes that something immoral ought to be so, based on the assumptions that the immoral thing happens and the fact that it ought to be the case that a certain agent knows that the thing happens. Åqvist [2014] writes that this paradox is “known under (...) names such as Åqvist’s Knower paradox and the “Knower””, but in the Stanford Encyclopedia of Philosophy this paradox is called the “Paradox of Epistemic Obligation” [McNamara, 2014]. The knower paradox considered in this thesis should not be confused with Fitch’s knowledge paradox or with Åqvist’s knower.

We discussed the definition of a paradox (Section 1.1), requirements on good solutions to paradoxes (Section 1.2), some ideas from epistemic logic (Section 1.3), and a few different formulations of the knower paradox (Section 1.4). In the last section of this chapter, we consider some recently published articles about this paradox.

1.5 The Current Debate on the Knower Paradox

There is only little consensus yet about how the knower paradox should be solved. It is generally accepted that the premise which says that knowledge implies truth is true, of which \(E1\) from Section 1.4.1 is an example and which corresponds to axiom scheme (A3) of epistemic logic. There are still debates going on about other parts of the paradox. Should the syntax be changed in such a way that statements that lead to paradoxes are abandoned? Should we accept the epistemic closure principle or not? We consider a few recent articles on these kinds of questions.

As we have seen in Section 1.4, Cross [2001] presents a version of the knower paradox in which the assumption of epistemic closure, \(E3\) by Kaplan...
and Montague, is dropped. An epistemic closure principle is “a principle stating that an agent’s knowledge is closed under some form of deductive consequence” [Dean and Kurokawa, 2014, p. 203]. Some examples of these closure principles are “if person $S$ knows $p$, and $p$ entails $q$, then $S$ knows $q$” and “$S$ knows $p$ and $p$ entails $q$ only if $S$ knows $q$” [Luper, 2012]. The principle $E3$, namely $[I(E_1, \neg D) \land K(E_1)] \rightarrow K(\neg D)$, is in line with this first example: ‘If $A$ knows $E_1$, and it is provable in the defined syntax that $E_1$ implies $\neg D$, then $A$ knows $\neg D$.’ If we generalize the statement by replacing $E_1$ by $p$ and $\neg D$ by $q$, then we have a very strong form of the principle. It implies that if $A$ knows at least one statement, then $A$ knows any theorem of the defined syntax (see [Stalnaker, 1991] for more on this problem of logical omniscience). This is the case because $I(\phi, \psi)$ is true for every theorem $\psi$. Some people (for example Maitzen [1998]) use the knower paradox as an argument against epistemic closure. Since Cross showed that leaving this premise out still results in a paradox, this argument does not hold.

Cross takes for granted that the generalization of $E1$, $K(\phi) \rightarrow \phi$, and the assumption $E1'$, $K'(\phi) \rightarrow \phi$, have the same status, but Uzquiano [2004] disagrees with him. Dean and Kurokawa [2014] write about the discussion between Cross and Uzquiano. In another recent article by Dean [2014], a paradox similar to the knower paradox is discussed, together with logic QLP, which contains both explicit modalities $t : \phi$ (“$t$ is a proof of $\phi$”) and proof quantifiers $(\exists x)x : \phi$ (“there exists a proof of $\phi$”).

Dean and Kurokawa also contribute to the question how the concept of knowledge should be defined by arguing that the knower paradox suggests that knowledge closely resembles mathematical provability. So the knower paradox, originally stated in 1960, is still a topical subject of research.

In addition to these articles, there is one by Paul Egré [2005], which we discuss at length in this thesis, and one by Poggiolesi [2007], which compares three different solutions to the knower paradox and comments on [Egré, 2005]. Egré argues that the knower paradox is solvable when modal provability logic is applied. He uses three different interpretations of this kind of logic to solve the paradox, namely interpretations by Skyrms [1978], Anderson [1983], and Solovay [1976].

In Chapter 2, we consider some ideas that are needed to understand these solutions and in Chapter 3 we state the solutions. In Chapter 4, we discuss the quality of Egré’s solutions. We check to what extent they satisfy Haack’s criteria, we evaluate whether some criticism on Egré’s article by Poggiolesi is valid, and we compare the comments by Dean and Kurokawa [2014] on a specific interpretation of provability logic with the corresponding solution by Egré. In this way, we explain to what extent the knower paradox can be solved using provability logic.
2 Provable Logic

In the previous chapter, the knower paradox is described, where a paradox is considered to be an apparently unacceptable conclusion derived by apparently acceptable reasoning from apparently acceptable premises [Sainsbury, 1987]. We intend to describe whether some solutions to the knower paradox, described by Paul Egré [2005], are satisfactory solutions. To understand these solutions, we need to know what provability logic is. George Boolos [1995, p. ix] states the following. “When modal logic is applied to the study of provability, it becomes provability logic”.

The most widely used provability logic is called GL and contains all axiom schemes from K (described in Section 1.3) and one extra scheme

\[ \Box(\Box \phi \rightarrow \phi) \rightarrow \Box \phi. \]  

(GL)

Like for epistemic logic, we have Kripke models \( \langle S, \pi, R_1, \ldots, R_m \rangle \) to define the semantics of provability logic. The logic K is valid in all Kripke frames. This is not the case for provability logic GL, since the theorems of GL are not exactly the sentences which are valid in all Kripke frames. To state for which subset of Kripke frames GL is valid, we need to define two notions. First we need \( R_i \) to be transitive, which means that \( (s_1, s_2) \in R_i \) if \( (s_1, s_3) \in R_i \) and \( (s_2, s_3) \in R_i \). Besides that, \( R_i \) should be conversely well-founded. A relation \( R_i \) is conversely well-founded if there does not exist an infinite sequence \( s_1, s_2, s_3, \ldots \) such that \( (s_1, s_2), (s_2, s_3), \ldots \in R_i \). The theorems of GL are exactly the sentences valid in all models \( \langle S, \pi, R_1, \ldots, R_m \rangle \) for which for all \( i = 1, \ldots, m \), \( R_i \) is transitive and conversely well-founded. The fact that we need the conversely well-foundedness of a Kripke model, implies that we need irreflexive models, since for a reflexive model we have an infinite sequence \( s_j, s_j, s_j, \ldots \) with \( (s_j, s_j), (s_j, s_j), \ldots \in R_i \).

We first consider some formal theories important for the study of provability, namely Robinson and Peano arithmetic and the relation between Peano arithmetic and modal logic. If a statement is proved in such a theory, then that statement is called a theorem. In addition to that, we look at some self-referential elements of arithmetic in Section 2.2. First Gödel numbering is considered, which allows us, for example, to formulate statements that say something about their own provability. Then we discuss the diagonal lemma, which proves that the self-referential statement from which we derived the knower paradox in Sections 1.1 and 1.4 can be defined in Peano arithmetic.

Kurt Gödel [1931/1986] proved two famous incompleteness theorems, which we discuss in Section 2.3. Martin Lüb [1955] proved a theorem about

\[9\]Instead of GL, this logic is sometimes called KW, KW, K4W, PrL or L.

\[10\]The axiom schemes in Section 1.3 contain ‘\( K_i \)’, which we replace by ‘\( \Box \)’ for the general variant of the schemes which can be used for different modal logics.
provability in Peano arithmetic and formulated some conditions on the provability predicate. We discuss these conditions and Löb’s theorem in Section 2.4, because together with Gödel’s insights these form the start of the development of provability logic [Verbrugge, 2014].

2.1 Formal Arithmetic

As Egré [2005, p. 16] says, “provability interpretations of modal logic are intrinsically related to formal arithmetic”. So, to understand the provability interpretations of modal logic which Egré applies to the knower paradox, we discuss formal arithmetic. We first describe the language of basic arithmetic and the definitions of Robinson and Peano arithmetic. These descriptions are based on respectively Section 4.3, Chapter 8, and Chapter 10 from [Smith, 2007]. In addition, the relation between PA and modal logic is explained.

2.1.1 The Language of Basic Arithmetic

The language of basic arithmetic is defined as an interpreted language $L_a = (L_A, \mathcal{I}_A)$. We discuss its syntax and its semantics. The syntax consists of logical and non-logical vocabulary of $L_A$, (standard) numerals, terms, a predicate, and well-formed formulae. The logical vocabulary of $L_A$ consists of connectives $[\neg, \land, \lor, \rightarrow, \leftrightarrow]$, brackets $[()]$, variables $[a, \ldots, d, u, \ldots, z$ and more], first-order quantifiers $[\forall, \exists]$, and the identity symbol $ [= ]$. The non-logical vocabulary of $L_A$ is $\{0, S, +, \cdot\}$. Here $0$ is a constant, $S$ is a one-place function called successor function, and $+$ and $\cdot$ are two-place functions. The numerals are expressions of the form $SS\ldots S0$, where ‘$S$’ occurs zero times or more. This general expression $SS\ldots S0$ can be abbreviated by ‘$n$’ if ‘$S$’ occurs exactly $n$ times, and $S0$ is abbreviated by 1, $SS0$ by 2, etcetera. In addition to numerals, the syntax contains terms. Any variable is a term and 0 is a term. Besides that, if $\sigma$ and $\tau$ are terms, then $S\sigma$, $(\sigma + \tau)$, and $(\sigma \cdot \tau)$ are terms too. The syntax does not contain any other terms than the ones described here. A term is called a closed term if it is variable-free, for example a numeral. There is only one predicate, namely the identity sign. The well-formed formulae are formed from atomic well-formed formulae by using connectives and quantifiers, where atomic well-formed formulae are of the form $\sigma = \tau$ for terms $\sigma$ and $\tau$. This completes the syntax of the interpreted language $L_a$. Now we discuss its semantics.

The interpretation $\mathcal{I}_A$ assigns values to terms in the following way. The value of ‘0’ is zero, which can be written down as $\text{val}[0] = 0$. If closed terms $\tau$ and $\sigma$ have values $\text{val}[\tau]$ and $\text{val}[\sigma]$ respectively, then $\text{val}[S\tau] = \text{val}[\tau] + 1$, $\text{val}[\sigma + \tau] = \text{val}[\sigma] + \text{val}[\tau]$, and $\text{val}[\sigma \cdot \tau] = \text{val}[\sigma] \times \text{val}[\tau]$. Besides this, $\sigma = \tau$ is true if and only if $\text{val}[\sigma] = \text{val}[\tau]$. A sentence of the form $\neg \phi$ is true if and only if $\phi$ is not true, and $(\phi \land \psi)$ is true if and only if both $\phi$ and $\psi$ are
true. Now only the quantifiers are still needed to be interpreted. Remember that \( n \) is the abbreviation of the general expression \( SS \ldots S0 \) where ‘\( S \)’ occurs \( n \) times. A sentence of the form \( \exists x \phi(x) \) is true if and only if, for some number \( n \), \( \phi(n) \) is true. Finally, a sentence of the form \( \forall x \phi(x) \) is true if and only if, for each number \( n \), \( \phi(n) \) is true. The domains of the functions \( \exists \) and \( \forall \) are the same, namely the set \( \mathbb{N} \) of natural numbers. We say that a well-formed sentence is true, if the statement which it expresses is true. So for example, the sentence \( \exists x \ 8 = (x+2) \) is true if there is some number such that eight is twice that number, so if eight is even.

Notice that we use sans-serif symbols for formal well-formed formulae, while for informal mathematics italic symbols are used\(^{11}\). Doing this, we follow the remarks by Smith [2007, Section 4.1].

This language \( L_A = \langle L_A, I_A \rangle \) of basic arithmetic is used for both Robinson and Peano arithmetic. In spite of the fact that Peano arithmetic (PA) was defined before Robinson arithmetic (Q), we discuss Q first, because PA is an extension of it.

### 2.1.2 Robinson Arithmetic

Robinson arithmetic is a formalized theory with language \( L_A \), a set of axioms, and a standard first-order arithmetic. It is abbreviated to Q and was first set out by Raphael M. Robinson [1950]. The axioms of Q are the following.

\[
\begin{align*}
\forall x (0 &\neq Sx) \quad \text{(1)} \\
\forall x \forall y (Sx = Sy &\rightarrow x = y) \quad \text{(2)} \\
\forall x (x &\neq 0 \rightarrow \exists y (x = Sy)) \quad \text{(3)} \\
\forall x (x+0 = x) &\quad \text{(4)} \\
\forall x \forall y (x+Sy = S(x+y)) &\quad \text{(5)} \\
\forall x (x \cdot 0 = 0) &\quad \text{(6)} \\
\forall x \forall y (x \cdot Sy = (x \cdot y) + x) &\quad \text{(7)}
\end{align*}
\]

The first axiom states that ‘0’ is not a successor of any numeral. Axiom (2) states that for any \( x \) and \( y \), the successors of \( x \) and \( y \) are identical only if \( x \) and \( y \) are identical themselves. As a third axiom, we see that for any nonzero \( x \), there is a \( y \) such that \( x = Sy \). In the description of Q by Boolos [1995, p. 49], this axiom is replaced by the equivalent axiom \( \forall x (x=0 \lor \exists y (x = Sy)) \). The fourth axiom states that \( x+0=x \) for every \( x \), but it appears to be impossible for Q to prove \( \forall x (0+x=x) \) [Smith, 2007, Section 8.4]. Axiom (5) states that

\(^{11}\)The two-place function ‘\( \cdot \)’ is used in formal well-formed formulae, while ‘\( \times \)’ is used for informal mathematics. There is no distinction between the different kinds of formulae for the symbols ‘\( + \)’, ‘\( = \)’, ‘\( \neg \)’, ‘\( \forall \)’, and ‘\( \exists \)’.
the value of $x$ added to the successor of $y$ is equal to the value of the successor of $x+y$ for each $x$ and each $y$. The statement that the product of any $x$ with 0 is equal to 0 is made by the sixth axiom, and the last axiom states that the product of any $x$ with the successor of any $y$ is equal to the addition of $x$ and the product of $x$ and $y$.

Together with a standard first-order arithmetic and the language $L_A$ which was discussed in Section 2.1.1, this set of axioms forms $Q$. Some statement $\phi$ is a theorem of $Q$ if it is (an instance of) an axiom or if it can be derived from the axioms. Statement $\phi$ can be derived from the axioms if there exists “a sequence $\phi_0, \ldots, \phi_n$ of [formulae from $L_A$] such that $\phi_n$ is $\phi$ and for each $i \leq n$, either $\phi_i$ is an axiom (...) or $\phi_i$ follows from some preceding members of the sequence using a rule of inference” [Hájek and Pudlák, 1993, p. 7–8]. The available rules of inference are modus ponens and generalization [Boolos, 1995, p. 19]. In Section 2.4.1 we state some conditions on derivability. If statement $\phi$ is a theorem of $Q$, this is denoted by ‘$\vdash \phi$’. We now consider an extension of Robinson arithmetic, namely Peano arithmetic.

### 2.1.3 Peano Arithmetic

Peano arithmetic (PA) is named after Giuseppe Peano [1889], who made a precise formulation of a set of axioms which was proposed by Richard Dedekind [1888]. To define PA, we need the following Induction Schema.

$\{ \phi(0) \land \forall x(\phi(x) \rightarrow \phi(Sx)) \} \rightarrow \forall x\phi(x)$  \hspace{1cm} (8)

According to Smith [2007, Section 10.4], “PA (…) is the first-order theory whose language is $L_A$ and whose axioms are those of $Q$ plus the universal closures of all instances of the Induction Schema”. So all axioms of $Q$, stated in Section 2.1.2, and each instance of this induction schema, are taken as axioms of PA.

Axiom (3) can be left out from the set of axioms of PA, because this one can be derived from an instance of the induction schema. We show how this is done by taking

$\phi(x) = (x \neq 0 \rightarrow \exists y (x=Sy))$.

Then we can substitute $x$ by 0. This results in $\phi(0) = (0 \neq 0 \rightarrow \exists y (0=Sy))$, which is true because the antecedent $0 \neq 0$ is false. So the first part, $\phi(0)$, of the antecedent of the induction instance is already true. We still need to consider the other part of the antecedent, $\forall x(\phi(x) \rightarrow \phi(Sx))$, where $\phi(Sx)$ is the formula $(Sx \neq 0 \rightarrow \exists y (Sx=Sy))$. This is true for all $x$, which we see if we substitute $x$ for $y$ in $(Sx=Sy)$. Since the consequent of $\phi(Sx)$ is true for all $x$, we conclude that $\phi(Sx)$ is true for all $x$. Since $\phi(Sx)$ is the consequent of $(\phi(x) \rightarrow \phi(Sx))$, we can conclude that $\forall x(\phi(x) \rightarrow \phi(Sx))$. This means
that both $\phi(0)$ and $\forall x (\phi(x) \to \phi(Sx))$ hold, so according to the induction schema it is the case that $\forall x (x \neq 0 \to \exists y (x = Sy))$ is a theorem of PA. Since $(x \neq 0 \to \exists y (x = Sy))$ is exactly what Axiom (3) of Q says, we conclude that this axiom can be left out from a definition of PA.

So just like Q, PA consists of a standard first-order arithmetic, the language $L_A$, and a set of seven axioms or axiom schemes. For PA, these are the induction schema (8) and all axioms from Q except for (3). In the same way as for Q, a statement of PA is (an instance of) an axiom (scheme) of PA or a well-formed formula derivable from these axioms. We denote the fact that $\phi$ is a theorem of PA by $\vdash \phi$. Since PA is an extension of Q, each theorem of Q is a theorem of PA.

Robinson arithmetic Q is sufficient to prove the diagonal lemma, which is discussed in Section 2.2.2. However, Q is not sufficient to develop for example the derivability conditions (see Section 2.4.1) to say something about its own syntax. Peano arithmetic is a theory which is able to prove results about its own syntax. According to Boolos [1995, p. 50], “among standard arithmetical theories capable of proving the diagonal lemma and results about their own syntax like the derivability conditions, PA is distinguished as the simplest, i.e. simplest to describe, now known”. Therefore, we use PA as the main arithmetical system for this thesis.

2.1.4 Peano Arithmetic and Modal Logic

This chapter is about provability logic, namely the application of modal logic to the study of provability. We have seen two theories, Robinson and Peano arithmetic, in which provability can be studied. What do these theories have to do with modal logic? We explain that there is a nice relation between a specific kind of modal logic and Peano arithmetic.

Modal logic is “the study of the deductive behavior of the expressions ‘it is necessary that’ and ‘it is possible that’” [Garson, 2014]. However, other systems, like epistemic logic, are also called modal logics. Usually ‘it is necessary that $\phi$’ is denoted by $\Box \phi$ and ‘it is possible that $\phi$’ is denoted by $\Diamond \phi$. In Section 1.3, we denote ‘it is known by agent $i$ that $\phi$’ by $K_i \phi$. When we consider provability logic, we denote ‘it is provable that $\phi$’ by $\Box \phi$. Remember we considered system $K$ with axiom schemes (A1) and (A2) on Page 12 and modus ponens and necessitation as rules of inference. As we stated at the beginning of this chapter, the best known system for provability logic is $GL$, which is system $K$ (with $K_i$ replaced by $\Box$) extended by axiom scheme $\Box(\Box \phi \to \phi) \to \Box \phi$. Solovay [1976, p. 289] adds $\Box \phi \to \Box \Box \phi$ as an axiom to $GL$. He states that this one can be derived from the other axioms, so it can be left out from a definition of $GL$. 

23
What is the relation between formal arithmetic and the system GL? On Page xxvi and further of [Boolos, 1995], we can read that a realization can be defined as a function that assigns a sentence of the language of arithmetic to a sentence letter of modal logic. For realizations * the following holds.

\[ \bot^* = \bot \]
\[ (\phi \rightarrow \psi)^* = (\phi^* \rightarrow \psi^*) \]
\[ (\square \phi)^* = \text{Prov}(\phi^*) \]

Other logical connectives like \((\phi \land \psi)\) can be defined by \(\rightarrow\) and \(\bot\), so * also respects these. The metalinguistic predicate \(\text{Prov}\) is sometimes called \(\text{Bew}\) after ‘beweisbar’. In Section 2.4.1 we discuss some important conditions on this predicate.

The system GL is arithmetically sound with respect to PA if GL ⊢ ϕ implies that for all realizations *, PA ⊢ ϕ*. According to Verbrugge [2014], “[i]t was already clear in the early seventies that GL is arithmetically sound with respect to PA”. So for any realization *, any theorem ϕ of GL can be translated to a theorem ϕ* of PA. The other way around also holds, so if PA ⊢ ϕ* for all realizations *, then GL ⊢ ϕ. This last statement, which states the arithmetical completeness of GL, has been proved by Robert Solovay [1976]. Since GL is arithmetically sound and arithmetically complete with respect to PA, it “prove[s] everything about the notion of provability that can be expressed in a propositional modal language and can be proved in Peano [a]rithmetic” [Verbrugge, 2014].

So the system GL provides the application of modal logic to the study of provability. We discuss some more about this study of provability in the rest of this chapter. Among the theorems that can be proved in arithmetic there are self-referential theorems. Thanks to Gödel numbering, we are able to formulate self-referential sentences within arithmetic (Section 2.2.1). This way of assigning ‘names’ to arithmetical statements is used in the so-called diagonal lemma (Section 2.2.2). After we state this lemma, we consider Gödel’s incompleteness theorems (Section 2.3) and another important theorem from provability logic together with some conditions for derivability (Section 2.4).

### 2.2 Self-Reference in Arithmetic

In Section 1.3 about epistemic logic, we consider the knowledge predicate \(K_i\), which is applied to sentences \(\phi\) as sentential operator. In Section 1.4, where three formulations of the knower paradox are described, the metalinguistic predicate \(K\) was used in order to apply the predicate to names \(\phi\) of sentences instead of sentences \(\phi\) itself. The provability predicate \(\text{Prov}(\phi)\) mentioned in Section 2.1.4 is also metalinguistic. We explain how names can be assigned
to sentences via Gödel numbering and state the diagonal lemma that proves the existence of self-referential sentences in Peano arithmetic.

### 2.2.1 Gödel Numbering

Using so-called Gödel numbering allows us to create self-referential statements in arithmetic. This is done by assigning numbers to expressions of the language $L_A$ of basic arithmetic (see Section 2.1.1). These numbers are called Gödel numbers and can be seen as the names of expressions.

Originally, Gödel [1931/1986, p. 179/157] used what we now call Gödel numbering in his proofs of his incompleteness theorems. We show an example of such a Gödel numbering, namely from Smith [2007, p. 125]. To define the Gödel number of a well-formed formula, we define code numbers for each symbol from $L_A$. Each symbol gets an odd number and variables get even numbers.

\[\neg \wedge \vee \rightarrow \leftrightarrow \forall \exists = ( ) 0 S + \cdot \]

\[\times \ y \ z \ \ldots \]

\[2 \ 4 \ 6 \ \ldots \]

**Definition** [Smith, 2007, p. 125] Let expression $e$ be the sequence of $k + 1$ symbols and/or variables $s_0, s_1, s_2, \ldots, s_k$. Then $e$’s Gödel number (g.n.) is calculated by taking the basic code number $c_i$ for each $s_i$ in turn, using $c_i$ as an exponent for the $i + 1$-th prime number $\pi_i$, and then multiplying the results, to get $\pi_0^{c_0} \cdot \pi_1^{c_1} \cdot \pi_2^{c_2} \ldots \cdot \pi_k^{c_k}$.

Like Egré [2005, p. 16], we denote the Gödel number of expression $e$ by $gn(e)$. We consider the well-formed formula $S0+S0=SS0$ corresponding to the sequence $s_0s_1s_2s_3s_4s_5s_6s_7s_8$. The symbols $s_0, s_3, s_6$, and $s_7$ are equal to ‘$S$’, so the exponents in the corresponding Gödel number are $c_0 = c_3 = c_6 = c_7 = 23$. In the same way we find $c_1 = c_4 = c_8 = 21$, $c_2 = 25$ and $c_5 = 15$. Then the Gödel number of $S0+S0=SS0$ is the following.

\[gn(S0+S0=SS0) = 2^{23} \cdot 3^{21} \cdot 5^{25} \cdot 7^{23} \cdot 11^{21} \cdot 13^{15} \cdot 17^{23} \cdot 19^{23} \cdot 23^{21}.\]

Defining Gödel numbers in this way, a formula has an unique Gödel number. The other way around, for a given Gödel number, the unique corresponding formula can be found via prime factorization according to the fundamental theorem of arithmetic [Hardy and Wright, 1938, Theorem 2].

We can also define *super Gödel numbers* [Smith, 2007, p. 127]. These are Gödel numbers of proofs, where a proof is considered as a sequence of well-formed formulae $e_0, e_1, \ldots, e_r$. Each formula $e_i$ ($i = 0, 1, \ldots, r$) gets its...
own Gödel number $g_i$ in the way defined above. Then we use multiplication of prime numbers again, and the super Gödel number of the proof becomes

$$2^{g_0} \cdot 3^{g_1} \cdot 5^{g_2} \ldots \cdot \pi_n^{g_n}.$$ 

Suppose we have the following proof of theorem $S_0 + S_0 = SS_0$ in PA, using expressions $e_0$ and $e_1$ and PA Axioms (4) and (5) from Section 2.1.2.

- $e_0 : S_0 + 0 = S_0$ by (4), $x = S_0$
- $e_1 : S_0 + S_0 = S(S_0 + 0)$ by (5), $x = S_0$, $y = 0$
- $t : S_0 + S_0 = SS_0$ by $e_0, e_1$

Then we have Gödel numbers

$$g_0 = 2^{23} \cdot 3^{21} \cdot 5^{25} \cdot 7^{21} \cdot 11^{15} \cdot 13^{23} \cdot 17^{21}$$
$$g_1 = 2^{23} \cdot 3^{21} \cdot 5^{25} \cdot 7^{23} \cdot 11^{21} \cdot 13^{15} \cdot 17^{23} \cdot 19^{17} \cdot 23^{23} \cdot 29^{21} \cdot 31^{25} \cdot 37^{21} \cdot 41^{19}$$
$$n = 2^{23} \cdot 3^{21} \cdot 5^{25} \cdot 7^{23} \cdot 11^{21} \cdot 13^{15} \cdot 17^{23} \cdot 19^{23} \cdot 23^{21}$$

of $e_0$, $e_1$ and $t$, respectively. This implies that the super Gödel number of the proof is

$$m = 2^{g_0} \cdot 3^{g_1} = 2^{23 \cdot 3^{21} \cdot 5^{25} \cdot 7^{23} \cdot 11^{15} \cdot 13^{23} \cdot 17^{21} \cdot 3^{23 \cdot 3^{21} \cdot 5^{25} \cdot 7^{23} \cdot 11^{21} \cdot 13^{15} \cdot 17^{23} \cdot 19^{17} \cdot 23^{23} \cdot 29^{21} \cdot 31^{25} \cdot 37^{21} \cdot 41^{19}}.$$ 

This Gödel numbering can be used to give names to sentences. The name of a sentence is the numeral denoting the Gödel number of that sentence.

So the name of expression $e_0$ is the numeral denoting the Gödel number $g_0$. This numeral is $SS \ldots S_0$, where ‘$S$’ occurs $g_0$ times. Let us call a numeral denoting a Gödel number a Gödel numeral. We denote the name of $e_0$ by $\overline{e_0}$, so $S_0 + 0 = S_0$ is the name of $S_0 + 0 = S_0$\footnote{Smith [2007] denotes the Gödel number of $\phi$ by $\overline{\phi}$ and the Gödel numeral of $\phi$ by $\overline{\overline{\phi}}$. Instead of following Smith, we follow Egré [2005] in denoting the Gödel number of $\phi$ by $gn(\phi)$ and Kaplan and Montague [1960] in denoting the numeral corresponding to this $gn(\phi)$ by $\overline{\phi}$.}. So $\overline{e_0}$ does not denote the Gödel number $g_0$ of $e_0$, but the Gödel numeral of $e_0$. Gödel numbering can be applied again to a numeral $\overline{e_0}$ denoting a Gödel number, but not to a Gödel number $g_0$ itself.

In Section 2.2.2, Gödel numbering is used in the formulation of the diagonal lemma. There we have a predicate which operates on names of sentences. Remember that predicates which operate on sentences, like they do in standard modal logic, are sentential, while if they operate on names (for example numerals denoting Gödel numbers) of sentences they are called metalinguistic. If in a certain system $\Sigma$, there exists a metalinguistic predicate $P$ and a sentence $\phi$ for which $\Sigma \vdash P(\overline{\phi}) \leftrightarrow \phi$, then we have a self-referential sentence $\phi$ which states about itself that it has property $P$. The diagonal lemma implies that in Peano arithmetic ($\Sigma = PA$), such sentences exist.
2.2.2 Diagonalization

In order to formulate some background for the solutions to the knower paradox described by Egré [2005], we discussed the language of arithmetic, Robinson and Peano arithmetic, the relation between Peano arithmetic and modal logic, and the idea of Gödel numbering. Egré discusses different ways to treat modalities and on Page 34, he defines ‘arithmetical treatment of modalities’ as “a treatment in which modalities are predicates applying to arithmetical codes of sentences, and in which any predicate can be diagonalized”. An example of these ‘arithmetical codes of sentences’ is the described Gödel numbering. What it means for a predicate to be ‘diagonalized’ is discussed here. In addition to that, the diagonal lemma is stated.

In the Stanford Encyclopedia of Philosophy, we see that diagonalization is a “general construction and proof method originally invented by Georg [Cantor, 1891] to prove the uncountability of the power set of the natural numbers” [Bolander, 2014]. According to Smith [2007, p. 130], the diagonalization of φ is the following substitution operation. Suppose we have a well-formed formula φ(y) with y as single free variable. This formula has Gödel number gn(φ), and we can substitute the numeral φ, denoting this Gödel number, for y. This results in the well-formed formula φ(φ).

We now consider the diagonal lemma\(^\text{14}\). This lemma proves that statement D, which is used to provide the original knower paradox in Section 1.4.1, can indeed be defined. Statement D satisfies Σ ⊢ D ↔ K(¬D).

The diagonal lemma is stated as follows.

**Theorem 2.1** (Diagonal Lemma, [Boolos, 1995, p. 54]\(^\text{15}\)). Suppose that P(y) is a formula of the language of PA in which no variable other than y is free. Then there exists a sentence S of the language of PA such that PA ⊢ S ↔ P(S).

A clear sketch of the proof can be found in a supplement of an article by Raatikainen [2014]. The sentence S for which PA ⊢ S ↔ P(S) holds, is comparable with a so-called fixed point a of function f for which f(a) = a holds. Therefore, the diagonal lemma is referred to as a fixed point theorem. However, this should not be confused with ‘the’ fixed point theorem for GL, which is independently proved by Dick de Jongh and Giovanni Sambin, and described by Boolos [1995, Chapter 8].

Since P(S) occurs in the theorem, P is a metalinguistic operator. This means that variable y of P(y) is the name of some formula, for example the numeral denoting the Gödel number of that formula. In the definition of statement D, which was used in the derivation of the original knower paradox, Smith [2007, p. 173] calls it ‘Diagonalization Lemma’ and explains that it deserves the status of being a theorem rather than being a lemma.

\(^{14}\)Smith [2007, p. 173] calls it ‘Diagonalization Lemma’ and explains that it deserves the status of being a theorem rather than being a lemma.

\(^{15}\)We replace Boolos’ “⌜S⌝” by ‘S’.

27
paradox (see Section 1.4.1), the predicate $P(y) = \neg K(y)$ is used. From the diagonal lemma it follows that there exists a sentence $S$, which we could call $\neg D$, satisfying $\text{PA} \vdash \neg D \leftrightarrow \neg K(\neg D)$. This implies that there exists a sentence $D$ such that $\text{PA} \vdash D \leftrightarrow K(\neg D)$, which is the first step in the derivation of the original knower paradox.

The diagonal lemma is used not only to show that certain sets of axioms lead to a contradiction like the knower paradox, but many theorems can be proved via this lemma. For example, the lemma appears in a proof of the arithmetical soundness of GL [Boolos, 1995, p. 60], and in a proof of Gödel’s first incompleteness theorem [Smith, 2007, p. 175].

We discussed Peano arithmetic and its relation to modal logic. We also considered the use of self-reference in arithmetic by explaining the concept of Gödel numbering and stating the diagonal lemma. In addition to this, we consider some other important theorems, namely Gödel’s incompleteness theorems and Löb’s theorem about provability in PA. Also the derivability conditions on the provability predicate are stated. All these elements are part of the logic of provability, which is relevant for the understanding of Egré’s solutions to the knower paradox.

2.3 Gödel’s Incompleteness Theorems

Kurt Gödel [1931/1986] proved two theorems about the incompleteness of a formal system which he calls $P$. This system, described in Section 2 of Gödel’s article, consists of the natural numbers, the axioms from Peano arithmetic (see Section 2.1.3), and the simple theory of types by Ramsey [1925]. One year after the famous incompleteness theorems were first published, Gödel [1932/1986] presented his theorems using Peano arithmetic without the extension of the simple theory of types [Kleene, 1986, p. 126]. In [Bernays and Hilbert, 1939], Gödel’s second incompleteness theorem is proved for Peano arithmetic\footnote{[Gödel, 1931/1986] contained only a sketch of the proof of this second theorem. Although Gödel intended to publish a second volume containing the formal proof, this never happened. According to Smoryński [1985, p. 3], this was “partly (I am told) due to Gödel’s health and partly (I am told) due to the immediate acceptance of his results”.

16}.

Before we can state Gödel’s incompleteness theorems, we need to give some definitions. A statement in the language of a system is called undecidable if it can be neither proved nor disproved in the system. The system is called incomplete if its language contains such an undecidable statement. As we saw in Section 1.2.1, a system $\Sigma$ is consistent if $\Sigma \vdash \bot$ does not hold. A system $\Sigma$ is $\omega$-consistent if there exists no well-formed formula $\phi(x)$ such that $\Sigma \vdash \exists x \phi(x)$ and $\Sigma \vdash \phi(a)$ for each formula $a$ representing a number. If some system is $\omega$-consistent, then it is also consistent, but there are con-
consistent systems which are \( \omega \)-inconsistent\(^{17} \). We state both incompleteness theorems for Peano arithmetic.

**Theorem 2.2** (Gödel’s First Incompleteness Theorem). *If PA is \( \omega \)-consistent, then PA is incomplete.*

John Barkley Rosser [1936] proved a stronger version of this theorem in which \( \omega \)-consistency was replaced by the weaker consistency. The theorem can be formulated differently by stating that *if PA is (\( \omega \)-)consistent, then any sentence asserting its own unprovability in PA is unprovable in PA*. A sentence asserting its own unprovability in PA is called a Gödel sentence. Such a sentence \( G \) can be formalized as follows, where \( Prov(\phi) \) is a provability predicate\(^{18} \) expressing the provability of \( \phi \) in PA.

\[
PA \vdash G \leftrightarrow \neg Prov(G)
\]

Apply the theorem to this sentence. First we suppose that PA is consistent. If sentence \( G \) would be provable, then \( G \) would be false by its definition. This would mean that PA contains a proof of a false sentence, which is in contradiction with the assumption that PA is consistent. We conclude that \( G \) is not provable in PA. With this we have arrived at the second stated formulation of Theorem 2.2. We can argue that this implies that PA is incomplete as follows. Since \( G \) is not provable in PA, \( G \) is true. This means that \( \neg G \) is false, and thus not provable in PA. So \( G \) is not disprovable. Since \( G \) is a sentence in PA which is neither provable nor disprovable in PA, it is undecidable in PA, which makes PA incomplete. Note that the conclusion that \( G \) is true does not lead to the paradox that \( G \) is both provable and not provable in PA, because the argumentation for the truth of \( G \) is not a proof in PA.

Is it possible to obtain a complete theory by adding \( G \) to PA as an axiom? Suppose we add \( G \) to PA, calling this new theory PA’. Then we can form another sentence \( G’ \), which satisfies \( PA’ \vdash G’ \leftrightarrow \neg Prov(G’) \). This sentence is undecidable in PA’ by the same argument as why \( G \) is undecidable in PA. We can keep adding these sentences to the theory, but we can always create a new undecidable Gödel sentence. So PA is not only incomplete, but also ‘incompletable’. Having discussed the first incompleteness theorem by Gödel, we consider his second incompleteness theorem too.

**Theorem 2.3** (Gödel’s Second Incompleteness Theorem). *If PA is consistent, then the consistency of PA cannot be proved within PA.*

We could also formulate this theorem as follows. If \( PA \nvdash \bot \), then it holds that \( PA \nvdash \neg Prov(\bot) \). We will give a proof of this theorem via Löb’s theorem.

\(^{17}\)A system is called (\( \omega \)-)inconsistent if it is not (\( \omega \)-)consistent. In the same way a system is complete if it is not incomplete.

\(^{18}\)Conditions on this predicate are stated in Section 2.4.1.
in Section 2.4.3. Is this incompleteness theorem very surprising? Suppose that the consistency of a theory could be proved within that theory. How reliable would this proof be? An inconsistent theory is also able to prove its own consistency, because everything follows from \( \bot \). So if we would have a theory which contains a theorem that expresses the consistency of that theory, then we would not be able to conclude whether that theory really is consistent. According to Gödel’s second incompleteness theorem, it follows that PA is inconsistent if the consistency of PA is provable in PA.

Since the incompleteness theorems are about theories proving properties of their own, they are important for the study of provability. There is another theorem, due to Martin Löb and Leon Henkin, which can also be added to this set of theorems which are important for provability logic. In addition to that, some conditions are needed to be satisfied by the provability predicate which we already referred to a few times. In the next section we see that a question from Henkin, inspired by Gödel’s first incompleteness theorem, led to Löb’s theorem and the derivability conditions.

2.4 Löb’s Theorem and Derivability Conditions

The diagonal lemma, discussed in Section 2.2.2, can be applied to a predicate \( \text{Prov}(y) \) which expresses that free variable \( y \) is the numeral denoting the Gödel number of a provable formula. In [Scholz et al., 1952], Leon Henkin asked a question that is nowadays usually stated in a way which is comparable to Gödel’s first incompleteness theorem, discussed in Section 2.3. Is \( H \), satisfying \( \text{PA} \vdash H \leftrightarrow \text{Prov}(H) \), provable or not? Martin Löb [1955] answered this question affirmatively. In the same article, Löb stated some conditions for provability predicates, which are now known as the derivability conditions. We discuss successively the derivability conditions, Henkin’s problem, and Löb’s solution to Henkin’s problem which results in Löb’s theorem.

2.4.1 Bernays-Hilbert-Löb Derivability Conditions

There are three conditions that a predicate \( \text{Prov}(\bar{\phi}) \) should satisfy in order to be a provability predicate for arithmetical theory \( \Sigma \).

\[
\begin{align*}
\text{L1} & \quad \Sigma \vdash \phi \Rightarrow \Sigma \vdash \text{Prov}(\bar{\phi}) \\
\text{L2} & \quad \Sigma \vdash \text{Prov}(\bar{\phi} \rightarrow \psi) \Rightarrow (\text{Prov}(\bar{\phi}) \rightarrow \text{Prov}(\bar{\psi})) \\
\text{L3} & \quad \Sigma \vdash \text{Prov}(\bar{\phi}) \Rightarrow \text{Prov}(\text{Prov}(\bar{\phi}))
\end{align*}
\]

These conditions are called the Hilbert-Bernays-Löb derivability conditions\(^{19}\) or just Löb’s derivability conditions\(^{20}\). A variant of these conditions was

\(^{19}\text{See } \text{Boolos, 1995, p. 16}, \text{ Smith, 2007, p. 223}.\)

\(^{20}\text{See } \text{Smoryński, 1991, p. 118}, \text{ Verbrugge, 2014}.\)
stated by Bernays and Hilbert [1939] for proving Gödel’s first incompleteness theorem. The currently accepted derivability conditions were first stated by Löb [1955] and are based on those of Bernays and Hilbert. The original ones by Bernays and Hilbert were not intended to define provability predicates, but those from Löb were.

The first derivability condition states that the statement that expresses that \( \phi \) is provable, is provable if statement \( \phi \) is provable itself. According to the second condition, it is provable that ‘the consequent of a conditional is provable if the conditional itself and the antecedent are both provable’. The last derivability condition states that it is provable that ‘if statement \( \phi \) is provable, then it is provable that statement \( \phi \) is provable’.

In the same article in which Löb proposed his variant of the derivability conditions, he answered a question posed by Henkin. In Section 2.4.2, we consider this question. We explain Löb’s answer in Section 2.4.3.

2.4.2 Henkin’s Problem

Leon Henkin posed a problem concerning provability. He formulated his question as follows.

If \( \Sigma \) is any standard formal system adequate for recursive number theory, a formula (having a certain integer \( q \) as its Gödel number) can be constructed which expresses the proposition that the formula with Gödel number \( q \) is provable in \( \Sigma \). Is this formula provable or independent in \( \Sigma \)? [Scholz et al., 1952]

We consider an arithmetical statement \( H \) that expresses its own provability in \( \Sigma \), where \( \Sigma \) is for example Peano arithmetic (see Section 2.1.3). The question is whether statement \( H \) is provable in \( \Sigma \) or not. If \( H \) is provable in \( \Sigma \), then \( H \) is true. If \( H \) is not provable in \( \Sigma \), then \( H \) is false. These conclusions are not helpful in finding out whether the statement is provable or not. Henkin asks whether the formula is provable or independent, where independent means undecidable, as defined in Section 2.3. Statement \( H \) is not refutable, which means that \( \Sigma \vdash \text{Prov}(\neg H) \) does not hold, for \( \text{Prov} \) denoting provability in \( \Sigma \). Suppose \( H \) would be refutable, then \( \Sigma \vdash \text{Prov}(\neg H) \) and thus by consistency of \( \Sigma \), \( \Sigma \vdash \neg \text{Prov}(H) \) would hold. This would imply that \( \Sigma \vdash \neg \text{Prov}(\neg H) \), i.e. \( \Sigma \) proves its own consistency. It would follow from Gödel’s second incompleteness theorem that the system is inconsistent, which we assume is not the case. Therefore, statement \( H \) is not refutable.

Georg Kreisel [1953], one of the authors of the problem section in which Henkin posted his problem, shows that the answer to Henkin’s question is both “yes” and “no”. He does this by constructing two predicates \( P_1 \) and \( P_2 \) expressing provability. For each predicate \( P_i \) he defines a sentence \( H_i \).
that expresses its own provability \((i = 1, 2)\). Sentence \(H_1\) appears to be \(P_1\)-provable and \(H_2\) appears not to be \(P_2\)-provable. The proof of the second sentence being not \(P_2\)-provable however, appears to solve the problem only if we do not require Gödel’s construction for the sentence that expresses its own provability (which we discussed in Section 2.3). According to Smoryński [1991, p. 114], Henkin’s question is not whether some sentence like \(0 = 0\) is accidentally equivalent to the expression of its own provability and he also did not ask whether all such sentences were provable, but Henkin “wanted to know if the (a?) sentence so constructed to express its own provability was provable”. According to Henkin [1954, p. 220] himself, Kreisel’s predicates \(P_1(a)\) and \(P_2(a)\) are formulae that do express provability, but do not express “the propositional function \(a\) is provable”. In addition, Henkin explains that “[t]he direct way to express \(a\) is provable is, of course, by the formula \((\exists x)B(x,a)\)”. This formula states that there exists some proof in \(\Sigma\) with Gödel numeral \(x\) of the statement with Gödel numeral \(a\). There are different ways to reformulate Henkin’s problem.

1. Formula \(H\) can be constructed which expresses the proposition that the formula with Gödel number \(\overline{H}\) is provable in \(\Sigma\). Is \(H\) provable? [Scholz et al., 1952]

2. Formula \(H\) can be constructed such that it expresses \((\exists x)B(x,\overline{H})\). Is \(H\) provable? [Henkin, 1954]

3. Suppose \(H\) satisfies \(\Sigma \vdash H \leftrightarrow \text{Prov}(\overline{H})\), where \(\text{Prov}(\overline{\phi})\) expresses provability of \(\phi\) in \(\Sigma\). Is \(H\) provable? [Boolos, 1995, Smith, 2007, p. 54, p. 229–230]

As opposed to formulations 1 and 2, where \(H\) is required to express its own provability, formulation 3 requires \(H\) only to be equivalent to its own provability. This means that formulation 3 is a generalization of formulation 1 as well as of 2. Kreisel solved the first formulation of Henkin’s problem, where he used that the sentence expressing its own provability does not necessarily need to be expressed via Gödel’s construction. The problem whether the formula \((\exists x)B(x,q)\), whose Gödel numeral is \(q\), is provable, is solved by Martin Löb [1955]. In the same article, he solves the more general formulation 3. In addition to that, he proves “a further generalization called Löb’s Theorem” [Smoryński, 1991, p. 117]. In Section 2.4.3, we consider this theorem and explain why it answers Henkin’s question.

\(^{21}\)We leave out the ‘or independent’ part of the question, because if the answer to the question ‘is \(H\) provable’ is ‘no’, it follows that \(H\) is independent.

\(^{22}\)Boolos [1995, p. 259, Footnote 1 of Chapter 11] realizes that Henkin’s questions is not equivalent to this reformulation.
2.4.3 L"ob’s Solution to Henkin’s Problem

Henkin’s problem as cited in Section 2.4.2 is answered both affirmatively and negatively by Kreisel [1953], where the negative answer did not require G"odel’s construction for the sentence that expresses its own provability. Both Henkin [1954, p. 220] and L"ob [1955, p. 115] state that Kreisel leaves open whether \((\exists x) B(x, q)\) with G"odel numeral \(q\) is provable or not. L"ob answers this question affirmatively by proving the generalization that \(H,\) satisfying \(\Sigma \vdash H \leftrightarrow \text{Prov}(H)\), is provable (for \(\text{Prov}(H)\) satisfying the derivability conditions from Section 2.4.1). The important ‘L"ob’s theorem’ followed from a proof in a previous version of L"ob’s article, as suggested by the referee. We use PA, described in Section 2.1.3, as theory \(\Sigma\).

**Theorem 2.4.** (L"ob’s Theorem) If \(PA \vdash \text{Prov}(S) \rightarrow S\), then \(PA \vdash S\).

This theorem can be proved as follows. We take \(\text{Prov}(y) \rightarrow S\) as formula of the language of PA and note that \(y\) is the only free variable of this formula. If we apply the diagonal lemma (see Section 2.2.2), it follows that there exists a sentence \(I\) such that \(PA \vdash I \leftrightarrow (\text{Prov}(I) \rightarrow S)\). In the formal proof that follows (L1), (L2), and (L3) refer to the Hilbert-Bernays-L"ob derivability conditions from Section 2.4.1, (R1) refers to modus ponens (see Section 1.3), and we use the notation of derivation rules by Meyer and van der Hoek [2004] again. For the sake of readability, we abbreviate \(\text{Prov}\) by \(P\).

\[
\begin{align*}
(1) & \quad PA \vdash P(\overline{S}) \rightarrow S & \text{by assumption} \\
(2) & \quad PA \vdash I \leftrightarrow (P(\overline{T}) \rightarrow S) & \text{by diagonal lemma} \\
(3) & \quad PA \vdash I \rightarrow (P(\overline{T}) \rightarrow S) & \text{by (2), EE} \\
(4) & \quad PA \vdash P(I \rightarrow (P(\overline{T}) \rightarrow S)) & \text{by (3), (L1)} \\
(5) & \quad PA \vdash P(I \rightarrow (P(\overline{T}) \rightarrow S)) \rightarrow (P(T) \rightarrow P(P(\overline{T}) \rightarrow S)) & \text{as instance of (L2)} \\
(6) & \quad PA \vdash P(T) \rightarrow P(P(\overline{T}) \rightarrow S) & \text{by (4), (5), (R1)} \\
(7) & \quad PA \vdash P(P(\overline{T}) \rightarrow S) \rightarrow (P(P(\overline{T})) \rightarrow P(S)) & \text{as instance of (L2)} \\
(8) & \quad PA \vdash P(\overline{T}) \rightarrow (P(P(\overline{T})) \rightarrow P(S)) & \text{by (6), (7), HS} \\
(9) & \quad PA \vdash P(\overline{T}) \rightarrow P(S) & \text{as instance of (L3)} \\
(10) & \quad PA \vdash P(T) \rightarrow P(S) & \text{by (8), (9), PC} \\
(11) & \quad PA \vdash P(\overline{T}) \rightarrow S & \text{by (10), (1), HS} \\
(12) & \quad PA \vdash (P(\overline{T}) \rightarrow S) \rightarrow I & \text{by (2), EE} \\
(13) & \quad PA \vdash I & \text{by (11), (12), (R1)} \\
(14) & \quad PA \vdash P(\overline{T}) & \text{by (13), (L1)} \\
(15) & \quad PA \vdash S & \text{by (14), (11), (R1)} 
\end{align*}
\]

\[23\] According to Smoryński [1991, p. 118], Halbach and Visser [2014, p. 8], this referee was Henkin.
The formalized version of this theorem is \( \text{PA} \vdash \text{Prov}(\text{Prov}(S) \rightarrow S) \rightarrow \text{Prov}(S) \). Note that this corresponds to the axiom scheme which was added to \( \text{K} \) to get \( \text{GL} \) in Section 2.1.4.

Why does this theorem solve Henkin’s problem, described in Section 2.4.2? Formulation 3, the generalization of the first two, was stated as the question whether \( H \), satisfying \( \Sigma \vdash H \leftrightarrow \text{Prov}(\overline{H}) \), is provable in \( \Sigma \). We take \( \Sigma = \text{PA} \). Since \( \text{PA} \vdash H \leftrightarrow \text{Prov}(\overline{H}) \), we have \( \text{PA} \vdash \text{Prov}(\overline{H}) \rightarrow H \), which implies \( \text{PA} \vdash H \), according to L¨ob’s theorem. This answers Henkin’s question affirmatively.

Consider the proof of L¨ob’s theorem. Since \( I \) is equivalent to \( \text{Prov}(\overline{I}) \rightarrow S \) according to Step (2), we have a construction comparable to Curry’s paradox, described in Section 1.1. L¨ob [1955, p. 117] mentions that the method used in his proof “leads to a new derivation of paradoxes in natural language” as an insight from the referee (namely Henkin). Maybe Curry himself has also mentioned the paradox to L¨ob, since Curry probably attended L¨ob’s presentation which he held before his article was published (see [L¨ob, 1954] for a summary), because he was the speaker just before L¨ob [Gerretsen and De Groot, 1954, p. 113].

In Section 2.3, G¨odel’s incompleteness theorems were considered. The second incompleteness theorem can be proved from L¨ob’s theorem. Remember that the incompleteness theorem can be formulated as follows. If \( \text{PA} \not\vdash \bot \), then \( \text{PA} \not\vdash \text{Prov}(\overline{\bot}) \). We prove this using the following variant of L¨ob’s theorem: If \( \text{PA} \not\vdash S \), then \( \text{PA} \not\vdash \text{Prov}(\overline{S}) \rightarrow S \). We substitute \( S \) by \( \bot \).

\[
\begin{align*}
(1) & \quad \text{PA} \not\vdash \bot \quad \text{by assumption} \\
(2) & \quad \text{PA} \not\vdash \text{Prov}(\overline{\bot}) \rightarrow \bot \quad \text{by L¨ob’s theorem} \\
(3) & \quad \text{PA} \not\vdash \neg \text{Prov}(\overline{\bot}) \quad \text{by PC}
\end{align*}
\]

If we consider the second incompleteness theorem for single-sentence extensions of \( \text{PA} \) instead of for \( \text{PA} \) itself, then L¨ob’s theorem can be inferred from it [Boolos, 1995, p. xxvi]. Suppose \( \text{PA}' \) is \( \text{PA} \) together with \( \neg S \). The consistency of \( \text{PA}' \) is expressed by \( \text{PA}' \vdash \neg \text{Prov}(\overline{S}) \). Then L¨ob’s theorem follows by the following proof.

\[
\begin{align*}
(1) & \quad \text{PA} \vdash \text{Prov}(\overline{S}) \rightarrow S \quad \text{by assumption} \\
(2) & \quad \text{PA} \vdash \neg S \rightarrow \neg \text{Prov}(\overline{S}) \quad \text{by (1), PC} \\
(3) & \quad \text{PA}' \vdash \neg \text{Prov}(\overline{S}) \quad \text{by (2), definition of \( \text{PA}' \)} \\
(4) & \quad \text{PA}' \not\vdash \neg \text{Prov}(\overline{S}) \quad \text{by (3), G¨odel’s second incompleteness theorem} \\
(5) & \quad \text{PA} \vdash S \quad \text{by (4), definition of \( \text{PA}' \)}
\end{align*}
\]

So G¨odel’s second incompleteness theorem can be proved via L¨ob’s theorem, and L¨ob’s theorem can be inferred from G¨odel’s theorem for some single-
To finish this chapter, we consider a list of important self-referencing statements defined in the first two chapters of this thesis.

\[ \Sigma \vdash D \leftrightarrow K(\neg D) \]
\[ \Sigma \vdash G \leftrightarrow \neg \text{Prov}(G) \]
\[ \Sigma \vdash H \leftrightarrow \text{Prov}(\perp) \]

Knower sentence \( D \) leads to the knower paradox, Gödel’s sentence \( G \) to Gödel’s first incompleteness theorem, and Henkin’s sentence \( H \) to Löb’s theorem. Why does \( D \) lead to a paradox, while changing the place of the negation sign to get \( G \) leads to a theorem? By the fixed point theorem by De Jongh and Sambin, \( \text{PA} \vdash G \leftrightarrow \neg \text{Prov}(G) \) is equivalent to \( \text{PA} \vdash G \leftrightarrow \neg \text{Prov}(\perp) \) (see [Boolos, 1995, p. xxx]). In addition, \( \text{PA} \vdash D \leftrightarrow \text{Prov}(\neg D) \) is equivalent to \( \text{PA} \vdash D \leftrightarrow \text{Prov}(\perp) \). So the Gödel sentence is equivalent to the assertion that the Gödel sentence is equivalent to the assertion that \( \text{PA} \) is consistent, while the knower sentence is equivalent to the assertion that the knower sentence is equivalent to the assertion that \( \text{PA} \) is inconsistent. This might be an indication for the difference between the paradox appearing from \( D \) and the theorem appearing from \( G \).

We try to solve the knower paradox using provability logic, in which the theorems by Gödel and Löb play an important role. As Visser [1998, p. 793] says, one advantage of provability logic is that “it gives us a direct way to compare notions such as knowledge with the notion of formal provability”. By interpreting knowledge as provability, some elements of the theory in which the knower paradox holds are rejected. If the resulting theory does not contain the knower paradox, then the paradox is solved. We consider some theories which solve the knower paradox according to Égré [2005] in Chapter 3. In Chapter 4, we discuss whether these solutions are satisfactory by discussing some articles that commented on them and by applying Haack’s requirements, described in Section 1.2.1, to the solutions.
3 Egré’s Solutions to the Knower Paradox

The aim of this thesis is to discuss to what extent provability logic can be used to solve the knower paradox. In Chapter 1, the knower paradox has been introduced both in an informal way as well as by three different formal formulations. The paradox consists of the apparently unacceptable conclusion that a certain apparently acceptable system of axioms and inference rules is inconsistent. This is shown by formulating the statement ‘we know that this statement is false’, or statement $D$ satisfying $\vdash D \leftrightarrow K(\neg D)$. In Chapter 2, the logic of provability has been described. Three examples of interpretations of provability logic, all discussed by Egré [2005], are formulated in the current chapter. In Chapter 4, we discuss whether these solutions are satisfactory by checking whether they satisfy Haack’s criteria, described in Section 1.2.1, and by discussing articles by Dean and Kurokawa [2014] and Poggiolesi [2007].

Before we consider the solutions Egré describes, we define four kinds of treatments of modalities, namely sentential, metalinguistic, syntactical and arithmetical treatments. Like in the first two chapters of this thesis, a sentential operator applies to sentences, but a metalinguistic predicate applies to names of sentences. If some metalinguistic predicate is self-referential, like a predicate to which the diagonal lemma from Section 2.2.2 applies, then we call it syntactical. Finally, an arithmetical predicate is a specific syntactical predicate, namely one which is self-referential because it can be diagonalized, and metalinguistic because it applies to arithmetical names of sentences. The relations between those different kinds of operators are shown in Figure 2.

![Figure 2: The relation between the different kinds of operators.](image)

Important in Egré’s article is that a syntactical treatment, by Montague [1963] and Cross [2001] defined without mentioning self-reference, is ambiguous between metalinguistic and self-referential treatment. When Montague states that a syntactical treatment of predicates is not possible without creating inconsistencies, he means a metalinguistic treatment which allows self-reference, as explained by Egré [2005, p. 34]. In addition, Egré shows
that there exists both a consistent metalinguistic treatment of modalities which is not self-referential and a consistent non-metalinguistic treatment which does allow self-reference.

Four different solutions to the knower paradox are mentioned by Egré [2005]. Some of them abandon self-reference, like Tarski’s hierarchy of languages, which is a non-satisfactory solution to the liar paradox (see Section 1.2.2). Others weaken the axiom schemes or the rules of inference. The first solution is the idea by Montague [1963] to abandon self-reference by re-treating from an arithmetical treatment of modalities to a standard modal treatment. This solution does not contain provability logic and we do not discuss it any further.

As a first interpretation of provability logic used to solve the knower paradox, Egré mentions a theory by Skyrms [1978] as a consistent metalinguistic treatment of modalities which does not allow self-referential statements like the knower sentence $D$ (satisfying $D \leftrightarrow K(\neg D)$). Instead of abandoning self-reference, Egré [2005, p. 38] prefers solutions in which self-reference is allowed, because a self-referential system of knowledge was used by Kaplan and Montague [1960] in order to deal with some paradox related to the knower paradox.

Two examples of self-referential systems are given, namely one by Anderson [1983], which weakens the axiom scheme $K(K(\overline{\phi}) \rightarrow \phi)$, and one by Solovay [1976], which weakens the necessitation rule of inference to prevent that scheme $K(K(\overline{\phi}) \rightarrow \phi)$ occurs by applying necessitation to the axiom scheme $K(\overline{\phi}) \rightarrow \phi$.

We discuss the systems by Skyrms, Anderson and Solovay respectively in Sections 3.1, 3.2 and 3.3. Although only Anderson published his system with the goal to contribute to the discussion about the knower paradox, Egré explains that all three of the theories provide solutions to the knower paradox. We consider the quality of these solutions in Chapter 4.

3.1 Skyrms’ Interpretation of Provability Logic

We consider Skyrms’ interpretation of provability logic and discuss why Egré states that this is a solution to the knower paradox. Skyrms [1978] himself does not mention the knower paradox in his article.

By the derivation of the original knower paradox in Section 1.4.1, we know that arithmetical treatment of modalities can lead to inconsistencies. Egré [2005] explains that Skyrms [1978] shows that there does exist a consistent form of metalinguistic treatment of modalities. Suppose $\Box \phi$ is metalin-

---

24We refer here to the unexpected examination paradox, from which the knower paradox arose (see [Kaplan and Montague, 1960]).
guistically interpreted as ‘φ is provable’. Skyrms defines modal language \( L_M \) as follows, where \( L_0 \) is a finitary language containing propositional calculus.

\[
L_M : \text{A language containing } L_0, \\
\text{closed under Boolean operators,} \\
\text{for which } \phi \in L_M \text{ implies } \Box \phi \in L_M
\]

The nested counterparts of \( L_M \) are based on the same language \( L_0 \).

\( L_0 \) : A finitary language containing propositional calculus.

\( L_{n+1} \) : A language based on \( L_0 \),

the smallest extension of \( L_n \) such that if \( \phi \in L_n \), then \( \text{Prov(‘} \phi \text{’)} \in L_{n+1} \),

closed under Boolean operators.

\[
L_\omega = \bigcup_{n \in \omega} L_n
\]

According to Skyrms [1978, p. 369], modal language \( L_M \) “may be thought of as the closure of \( L_0 \) under modalities as sentential operators and truth functions”. The predicate \( \text{Prov(‘} \phi \text{’)} \) expresses that \( \phi \) is provable, where the quotes are symbols of the object-language. Skyrms uses \( *Q(S) \), where the asterisk is the part interpreted as ‘is valid’ or ‘is provable’. Egré does not discuss the interpretation as validity, but only considers the provability interpretation and writes \( *Q(S) \) as \( \text{Prov(‘} \phi \text{’)} \). Skyrms [1978, p. 369–370] explains that “[t]he expression consisting of a sentence prefixed by ‘Q’ is to be thought of as a name for that sentence”. So ‘\( \phi \)’ in \( \text{Prov(‘} \phi \text{’)} \) does not express the numeral corresponding to the name of \( \phi \), but expresses just the name of \( \phi \). This means that the treatment of modalities in Skyrms’ system is metalinguistic, because the modalities are applied to names of sentences and not to sentences themselves. It is not syntactical, since it does not allow self-reference\(^{25}\).

The modal language \( L_M \) needs to be translated to metalanguage \( L_\omega \). Each sentence of \( L_M \) gets assigned a metalinguistic correlate in \( L_\omega \) via the translation \( t : L_M \to L_\omega \), which satisfies the following criteria.

\[
t(\phi) = \phi \quad \text{for all } \phi \in L_0 \\
t(\Box \phi) = \text{Prov(‘} t(\phi) \text{’)} \quad \text{for all } \phi \in L_M \\
t \text{ distributes over the truth-functional connectives}
\]

Using this translation, the modal degree of \( \phi \in L_M \) gives the index of the first language to which \( t(\phi) \) belongs.

\(^{25}\) According to Egré [2005, p. 37], modalities are treated syntactically in Skyrms’ system. This is because in this part of his analysis of syntactical treatment of modalities, Egré considers syntactical predicates as predicates taking names or codes of sentences as arguments. On Page 36, we defined syntactical predicates in a stronger sense, namely as metalinguistic predicates with self-reference.
Why is Skyrms’ system, consisting of an hierarchy of languages and a translation from $L_M$ to $L_\omega$, a solution to the knower paradox? Egré states the following consistency result.

**Theorem 3.1** (Consistency of Skyrms’ System [Egré, 2005, p. 36–37]). Let $L_0$ be the language of Robinson arithmetic $Q$. Let $T_0 = Q$ and consider the chain of (deductively closed) theories $T_n$ in the languages $L_n$ previously specified, where $T_{n+1}$ is the smallest extension of $T_n$ satisfying:

1. If $\phi \in T_n$, then $\text{Prov}(\langle \phi \rangle) \in T_{n+1}$
2. If $\phi, \psi \in L_n$, then $\text{Prov}(\langle \phi \rightarrow \psi \rangle) \rightarrow (\text{Prov}(\langle \phi \rangle) \rightarrow \text{Prov}(\langle \psi \rangle)) \in T_{n+1}$
3. If $\phi \in L_n$, then $\text{Prov}(\langle \phi \rangle) \rightarrow \phi \in T_{n+1}$
4. If $\text{Prov}(\langle \phi \rangle) \in L_n$, then $\text{Prov}(\langle \phi \rangle) \rightarrow \text{Prov}(\langle \text{Prov}(\langle \phi \rangle) \rangle) \in T_{n+1}$

The theory $T_\omega = \bigcup_{n \in \omega} T_n$ is consistent if $Q$ is consistent.

According to Egré [2005, p. 37], who gives a short proof of this theorem, “[t]his consistency result shows that the theory $T_\omega$, although it is an extension of Robinson arithmetic, can satisfy all the metalinguistic translations of the modal schemata involved in (...) the weak system T-Nec used to present the [knower paradox]”. The way in which Egré [2005, p. 18] presents the knower paradox using what he calls system T-Nec is via a theorem about theory $T$ extending $Q$, and $K$ as unary predicate of the language of $T$ satisfying $T \vdash K(\phi) \rightarrow \phi$ (scheme (A3) from Section 1.3) and the scheme ‘if $T \vdash \phi$, then $T \vdash K(\phi)$’ (necessitation rule (R2) from Section 1.3). It follows that $T$ is inconsistent. This is a version of the knower paradox in which the unacceptable conclusion is ‘this system representing knowledge is inconsistent’. Like T-Nec, Skyrms’ system is an extension of $Q$ and represents knowledge satisfying (A3) and (R2), but in Skyrms’ system, the predicate Prov is treated in a way that does not allow self-referential sentences like $D \leftrightarrow \text{Prov}(\langle \neg D \rangle)$. This difference leads to the difference in consistency of the systems.

Why is $D \leftrightarrow K(\overline{D})$ not allowed in Skyrms’ system, meaning that there is no $D \in L_\omega$ satisfying $D \leftrightarrow \text{Prov}(\langle \neg D \rangle) \in T_\omega$? It is not the case that every form of self-reference is disallowed. “[T]here is self-referential apparatus in the theories considered by Skyrms, because these extend weak arithmetic, but this form of self-reference cannot interfere with the metalinguistic predicate Prov that [Skyrms] considers” [Egré, 2005, p. 37]. Egré [2005, p. 37–38] also states that “[t]he core of Skyrms’ approach is indeed to sever the self-referential apparatus of arithmetic from the metalinguistic system used to handle the predicate Prov”. It is not clear whether Egré suggests that the knower sentence $D$, with $D \leftrightarrow \text{Prov}(\langle \neg D \rangle)$, is not in Skyrms’
language $L_\omega$ or not in his theory $T_\omega$. We notice that $D \leftrightarrow \text{Prov}(\neg D')$ does occur in the language $L_\omega$, but not in theory $T_\omega$. We first explain why $D \leftrightarrow \text{Prov}(\neg D') \in L_\omega$.

Suppose $D \in L_n$, then $\neg D \in L_n$, because $L_n$ is closed under Boolean operators. Then $\text{Prov}(\neg D') \in L_{n+1}$ follows by definition of $L_{n+1}$. Since $D \in L_n$, $D \in L_{n+1}$ follows. Now we know that $D, \text{Prov}(\neg D') \in L_{n+1}$. $L_{n+1}$ is closed under Boolean operators, so $D \leftrightarrow \text{Prov}(\neg D') \in L_{n+1}$. So as long as $D \in L_n$, $D \leftrightarrow \text{Prov}(\neg D') \in L_{n+1}$ holds. $L_n$ is not empty, because $L_0$ is not empty, so there exists a $D \in L_n$ to form the knower sentence in $L_{n+1}$ for certain $n$. Since $L_{n+1} \subset L_\omega$, there exists a $D$ such that $D \leftrightarrow \text{Prov}(\neg D') \in L_\omega$.

In addition, we prove that $D \leftrightarrow \text{Prov}(\neg D') \notin T_\omega$ by contradiction. Suppose $D \leftrightarrow \text{Prov}(\neg D') \in T_\omega = \cup_{n \in \omega} T_n$, for some $D \in L_\omega$. Then there exists some $n \in \omega$ such that $D \leftrightarrow \text{Prov}(\neg D') \in T_n$. This means that for this $n \in \omega$ either $D, \text{Prov}(\neg D') \in T_n$, or $\neg D, \neg \text{Prov}(\neg D') \in T_n$. We consider both situations.

Suppose there exists an $n \in \omega$ such that $D, \text{Prov}(\neg D') \in T_n$. Something can only be in $T_n$ if it is also in $L_n$. So $\text{Prov}(\neg D') \in L_n$. According to the definition of $L_n$, it follows that $\neg D \in L_{n-1}$.

By the third one of the four requirements in Theorem 3.1, it follows that $\text{Prov}(\neg D') \rightarrow \neg D \in T_n$. Since $\text{Prov}(\neg D') \in T_n$ and $T_n$ is deductively closed, we conclude that $\neg D \in T_n$. We already knew that $D \in T_n$, and $T_n$ is consistent, so there is a contradiction. This means that there does not exist an $n \in \omega$ such that $D, \text{Prov}(\neg D') \in T_n$.

Now we consider the second situation. Suppose there exists an $n \in \omega$ such that $\neg D, \neg \text{Prov}(\neg D') \in T_n$. Because $\neg D \in T_n$, it follows by the first one of the four requirements in Theorem 3.1 that $\text{Prov}(\neg D') \in T_{n+1}$. We also have $\neg \text{Prov}(\neg D') \in T_{n+1}$, because $T_{n+1}$ is an extension of $T_n$, and $\neg \text{Prov}(\neg D') \in T_n$. Because we have $\text{Prov}(\neg D'), \neg \text{Prov}(\neg D') \in T_{n+1}$ and $T_{n+1}$ is consistent, there is a contradiction. So there is no $n \in \omega$ such that $\neg D, \neg \text{Prov}(\neg D') \in T_n$. Considering both possible situations, we conclude that there is no $n \in \omega$ such that $D, \text{Prov}(\neg D') \in T_n$ or $\neg D, \neg \text{Prov}(\neg D') \in T_n$. It follows that there does not exist an $n \in \omega$ such that $D \leftrightarrow \text{Prov}(\neg D') \in T_n$, so we conclude that $D \leftrightarrow \text{Prov}(\neg D') \notin T_\omega$.

The first step of the original derivation of the paradox consisted of $\vdash D \leftrightarrow K(\neg D')$ (see Section 1.4.1, Page 14). Since $D \leftrightarrow \text{Prov}(\neg D') \notin T_\omega$, so $T_\omega \not\vdash D \leftrightarrow \text{Prov}(\neg D')$, the knower paradox cannot be derived in Skyrms’ system $T_\omega$ in the same way as we did in Section 1.4.1. There is not any way to derive inconsistency from $\vdash D \leftrightarrow \text{Prov}(\neg D')$ in a system for which $\not\vdash D \leftrightarrow \text{Prov}(\neg D')$, so there is not any way to derive the knower paradox

\footnote{The case $n = 0$ is an exception. Since $\text{Prov}(\neg \phi) \notin L_0$ for arbitrary $\phi$, $\text{Prov}(\neg D') \notin T_0$ holds. So it immediately follows that $D, \text{Prov}(\neg D') \notin T_0$.}
in $T_\omega$. Therefore, accepting $T_\omega$ solves the knower paradox. In Section 4.1 we consider the quality of this solution, but first we introduce two other provability interpretations as solutions to the knower paradox.

### 3.2 Anderson’s Interpretation of Provability Logic

In addition to Skyrms’ interpretation of provability logic, we consider Anderson’s provability interpretation of epistemic logic. We discuss why both Anderson and Egré state that this is a solution to the knower paradox.

Skyrms’ system as a solution to the knower paradox abandoned a certain form of self-reference in his theory $T_\omega$. Anderson [1983] argues that we should not abandon self-reference, but modify the incompatible axiom schemes which lead to the paradox. Anderson considers the three generalizations of the axioms $E1$, $E2$, and $E3$ from the original knower paradox by Kaplan and Montague [1960] (see Section 1.4.1). Like Egré, we call these schemes $T$, $U$, and $I$.

\[
K(\bar{\phi}) \rightarrow \phi \quad (T)
\]
\[
K(K(\bar{\phi}) \rightarrow \phi) \quad (U)
\]
\[
[I(\bar{\phi}, \bar{\psi}) \land K(\bar{\phi})] \rightarrow K(\bar{\psi}) \quad (I)
\]

As we will see, Anderson constructs a hierarchy which allows self-reference in a way in which $(T)$ and $(I)$ still hold, but $(U)$ is not valid anymore. His hierarchy of languages is defined as follows\(^{27}\), where $L_A$ is the language of Robinson and Peano arithmetic (see Section 2.1.1).

\[
L_0 : \text{the smallest extension of } L_A \text{ such that}
\]
\[
\text{if } \phi, \psi \in L_A, \text{ then } K_0(\bar{\phi}), I_0(\bar{\phi}, \bar{\psi}) \in L_0, \text{ closed under Boolean operators.}
\]

\[
L_{i+1} : \text{the smallest extension of } L_i \text{ such that}
\]
\[
\text{if } \phi, \psi \in L_i, \text{ then } K_{i+1}(\bar{\phi}), I_{i+1}(\bar{\phi}, \bar{\psi}) \in L_{i+1}, \text{ closed under Boolean operators.}
\]

\[
L_\omega = \bigcup_{i \in \omega} L_i
\]

Notice that this $K_i$ does not mean ‘agent $i$ knows’, like it did in Section 1.3, but it indicates a certain level of knowledge. Anderson [1983, p. 348–349] gives an “intuitive motivation” for accepting more than one knowledge predicate. If within some system $Q'$, based on Robinson arithmetic $Q$, the knowledge $K_0$ of some person which we call Oscar is represented, then there are

\(^{27}\)Egré [2005, p. 39] defines $L_{i+1}$ as $L_i \cup \{K_i, I_i\}$, which implies that $L_1 = L_0 \cup \{K_0, I_0\} = L_0$. Anderson [1983, p. 351–352] himself states that “language $L_i$ is obtained from $L_\omega$ by omitting all $K$ and $I$ predicates with subscripts greater than $i$”. So instead of adding $K_i$ and $I_i$ in language $L_{i+1}$, we add $K_{i+1}$ and $I_{i+1}$.
provable statements which cannot be in \( Q' \). Examples of this are the consistency of \( Q' \) and the Gödel sentence \( G \), with \( G \leftrightarrow \neg \text{Prov}(G) \), where \( \text{Prov} \) means provability in \( Q' \). After reflecting on some arguments, Oscar can understand the proofs of these statements that cannot be in \( Q' \), the system representing Oscar’s knowledge. Therefore, it seems intuitive to accept different knowledge levels representing different parts of Oscar’s knowledge.

The lowest level is \( K_0 \), and knowledge level \( K_1 \) is described by the theorems of \( K_0 \) together with the statements which could not be in \( Q' \) but became known by Oscar. In the same way we can argue the existence of some knowledge level \( K_{i+1} \) for \( i \geq 1 \), because there are provable, and thus knowable, statements which cannot be contained in \( K_i \).

It is assumed that there is a given Gödel numbering for \( L_\omega \), and we define \( gn(L_\omega) = \{ gn(l) \mid l \in L_\omega \} \). Then the semantics of Anderson’s hierarchy of languages is as follows, where \( V_p \) is an interpretation of \( L_A \) on which a chain of interpretations \( V_i \) is based.

\[
\begin{align*}
V_0 & \text{ extends } V_p \text{ to } L_0 \\
V_{i+1} & \text{ extends } V_i \text{ to } L_{i+1} \\
V_i(K_i) & \subseteq gn(L_\omega) \\
V_i(I_i) & \subseteq gn(L_\omega) \times gn(L_\omega) \\
V & = \bigcup_{i \in \omega} V_i
\end{align*}
\]

The hierarchy of theories with sequence of axiom sets \( (T_i)_{i \in \omega} \) and sequence of interpretations \( (V_i)_{i \in \omega} \) are defined as follows.

\[
\begin{align*}
T_0 &= Q \cup \{ K_0(\overline{\phi}) \rightarrow \phi \mid \phi \in L_\omega \} \\
T_{i+1} &= T_i \cup \{ K_{i+1}(\overline{\phi}) \rightarrow \phi \mid \phi \in L_\omega \} \\
V_0(K_0(\overline{\phi})) &= 1 \text{ if and only if } Q \vdash \phi \\
V_{i+1}(K_{i+1}(\overline{\phi})) &= 1 \text{ if and only if } T_i \vdash \phi \\
V_0(I_0(\overline{\phi}, \overline{\psi})) &= 1 \text{ if and only if } Q \vdash \phi \rightarrow \psi \\
V_{i+1}(I_{i+1}(\overline{\phi}, \overline{\psi})) &= 1 \text{ if and only if } T_i \vdash \phi \rightarrow \psi
\end{align*}
\]

In this thesis, we consider axiom set \( T_\omega = \bigcup_{i \in \omega} T_i \) as Anderson’s theory or Anderson’s system. Anderson’s sequence of provability interpretations of knowledge is coherent, which means that the following constraints are

---

28 Suppose that \( G \) is provable in \( Q' \). By the same reasoning by which we concluded that \( G \) is not provable in PA (on Page 29 in Section 2.3), \( G \) is not provable in \( Q' \). So \( G \) is not in \( K_0 \), but we (at least Gödel) can give a proof of the fact that \( G \) is not provable in \( Q' \). By using the definition of \( G \), we prove that \( G \) is true.

29 Anderson [1983, p. 348–349] could have paid more attention to some details of his argument. We discuss this in Section 4.2.2.

42
satisfied for all levels $i, j$.

$$V_i(K_i) \subseteq V_{i+1}(K_{i+1})$$

$$V_i(I_i) \subseteq V_{i+1}(I_{i+1})$$

If $n = gn(\phi) \in V_i(K_i)$, then $\exists j \geq i$ such that $V_j(\phi) = 1$.

If $n = gn(\phi), m = gn(\psi), (n, m) \in V_i(I_i)$, then $\exists j \geq i$ such that $V_j(\phi \rightarrow \psi) = 1$.

If $(n, m) \in V_i(I_i), n \in V_i(K_i)$, then $m \in V_i(K_i)$.

In addition to the fact that the sequence of interpretations is coherent, the following statements are satisfied for all levels $i$.

$$V(K_i(\phi) \rightarrow \phi) = 1$$

$$V([I_i(\phi, \psi) \land K_i(\phi)] \rightarrow K_i(\psi)) = 1$$

$$V(K_{i+1}(K_i(\phi) \rightarrow \phi)) = 1$$

By the first two of those statements, we still have $(T)$ and $(I)$ in Anderson’s system. There are two different forms of $(U)$, namely $K_{i+1}(K_i(\phi) \rightarrow \phi)$ and $K_i(K_i(\phi) \rightarrow \phi)$. The first one is valid, but if we use this one in the derivation of the knower paradox as described in Section 1.4.1 on Page 14, then we will not arrive at an inconsistency. This is the case, because we get $K_{i+1}(\neg D)$ in Step (10) of the derivation and $K_i(\neg D) \rightarrow D$ in Step (11), which does not give us $D$. Therefore, we cannot conclude the inconsistency of $D$ with $\neg D$. Applying the other form, $K_i(K_i(\phi) \rightarrow \phi)$, would lead to the inconsistency in the same way as described in Section 1.4.1 by replacing $K$ with $K_i$. However, this form of $U$ is not valid in Anderson’s system, because by definition of theory $T_j, T_j \vdash K_i(\phi) \rightarrow \phi$ holds only for $j \geq i$. This means that $T_{i-1} \vdash K_i(\phi) \rightarrow \phi$ does not hold, so by definition of interpretation $V_i, V_i(K_i(\phi) \rightarrow \phi) \neq 1$. So this second form of $U$ is not valid.

Since this second form of $U$, which would lead to the knower paradox, is not valid, the paradox is solved by Anderson’s provability interpretation. We will see whether this solution is satisfactory in Chapter 4. Egré [2005, p. 40] states that “[t]he strength of this solution, as compared to [Skyrms’ system], is to license the construction of self-referential statements at every level of the hierarchy”. We consider another system of modal logic, by Solovay [1976], as solution to the knower paradox, which according to Egré [2005, p. 38] has a “significant connection” with Anderson’s system.

### 3.3 Solovay’s Interpretation of Provability Logic

We discussed Skyrms’ consistent system, in which there is one provability predicate and self-referential sentences cannot be proved in $T_\omega$. We also considered Anderson’s hierarchy of languages, in which infinitely many provability predicates occur but self-referential sentences can be valid. According
to Egré [2005, p. 40], the framework of modal provability logic combines the possibility of self-reference with the use of only one provability predicate. Like Skyrms, Solovay did not publish his theory in the context of the knower paradox.

Remember that the system GL contains the propositional tautologies as axioms as well as all instances of the schemes \((\Box \phi \land \Box (\phi \rightarrow \psi)) \rightarrow \Box \psi\) and \(\Box (\Box \phi \rightarrow \phi) \rightarrow \Box \phi\), and the inference rules modus ponens and necessitation (see Section 2.1.4). The system GLS, defined by Solovay [1976, Section 5.1]\(^\text{30}\), contains all theorems of GL as axioms as well as all instances of the reflection principle \(\Box \phi \rightarrow \phi\), and modus ponens is its single rule of inference. Like for GL, the arithmetical soundness and the arithmetical completeness of the system GLS can be proved, but with respect to the standard model \(\langle \omega; +, \cdot \rangle\) instead of to PA [Solovay, 1976].

Why is the knower paradox prevented in GLS? Remember that \(K(E1)\) was needed in the derivation of the knower paradox by Kaplan and Montague [1960] (see Section 1.4.1, Page 14, Step (7)), where \(E1\) was defined as \(K(\neg D) \rightarrow \neg D\). In GLS, we have \(\Box \neg D \rightarrow \neg D\) as an instance of the reflection principle. Because necessitation is not an inference rule of GLS, \(\Box (\Box \neg D \rightarrow \neg D)\) cannot be derived from the reflection principle here. This means that the derivation by Kaplan and Montague cannot be repeated in GLS.

Egré states more about Solovay’s system, in particular about its connection to the one by Anderson. We cite\(^\text{31}\) him and give some comments on it.

The system GLS corresponds to the system \(PA^+\) obtained by closure under modus ponens from PA supplemented with all instances of the reflection principle. \(PA^+\) is stronger than PA because it can now prove the consistency of PA; \(PA^+\) is therefore the counterpart of the first system \(T_0\) in Anderson’s progression. What this shows however is what remained only hinted at in Anderson’s treatment, namely the fact that when knowledge is interpreted in terms of provability, an implicit hierarchy is present within the first stage of the progression: in order to keep principle \((T)\), one needs to restrict the rules of inference governing its interaction with [necessitation]. [Egré, 2005, p. 43]

Egré introduces a set \(PA^+\) of theorems as an extension of the set of theorems of PA. In addition to the theorems of PA, \(PA^+\) contains all instances of \(\text{Prov}(\overline{\phi}) \rightarrow \phi\), where \(\text{Prov}\) means provability in PA, and \(PA^+\) is closed under modus ponens but not under necessitation. Egré makes the following

\(^{30}\)Solovay’s \(G\) is our GL and his \(G’\) is our GLS.

\(^{31}\)We use our own notation of GLS, PA, T, etc.
four claims. The first is that \textit{GLS} corresponds to \(\text{PA}^+\). The second claim is that \(\text{PA}^+\) can prove the consistency of \(\text{PA}\). As the last two claims, Egré states that \(\text{PA}^+\) is the counterpart of \(T_0\) in Anderson’s system (described in Section 3.2) and that \(T_0\) contains an implicit hierarchy. We discuss these four claims consecutively.

\textbf{(1) Why does GLS correspond to \(\text{PA}^+\)?} We think Egré means that \(\text{GLS}\) corresponds to \(\text{PA}^+\) in the same way as \(\text{GL}\) corresponds to \(\text{PA}\). The system \(\text{GL}\) is arithmetically sound and arithmetically complete with respect to \(\text{PA}\), which means that \(\text{GL} \vdash \phi\) if and only if \(\text{PA} \vdash \phi^*\) for all realizations \(*\). Is it the case that \(\text{GLS} \vdash \phi\) if and only if \(\text{PA}^+ \vdash \phi^*\) for all realizations \(*\), so we can say that \(\text{GLS}\) corresponds to \(\text{PA}^+\)?

Solovay [1976, Section 5.1] proves that \(\text{GLS}\) is arithmetically sound and arithmetically complete with respect to the standard model \(\langle \omega; +, \cdot \rangle\). In addition, \(\text{PA}^+\) is sound with respect to this standard model, so \(\text{PA}^+ \vdash \phi\) implies \(\omega \models \phi\). So if \(\text{GLS} \not\vdash \phi\), then by the completeness part of Solovay’s theorem some realization \(*\) exists such that \(\omega \not\models \phi^*\). By soundness of \(\text{PA}^+\) with respect to \(\omega\), it follows that \(\text{PA}^+ \not\vdash \phi^*\). So assuming \(\text{GLS} \not\vdash \phi\), it follows that \(\text{PA}^+ \not\vdash \phi^*\) for some realization \(*\). This means that \(\text{GLS}\) is arithmetically complete with respect to \(\text{PA}^+\).

To prove the arithmetical soundness of \(\text{GLS}\) with respect to \(\text{PA}^+\), the arithmetical soundness of \(\text{GL}\) with respect to \(\text{PA}\) can be used. We also use a theorem from Boolos [1995, p. 131], according to which \(\text{GLS} \vdash \phi\) implies that there exist \(\psi_1, \ldots, \psi_n\) such that \(\text{GL} \vdash \{\Box \psi_i \rightarrow \psi_i \mid i = 1, \ldots, n\} \rightarrow \phi\). Since \(\text{GL}\) is sound with respect to \(\text{PA}\), it follows that \(\text{PA} \vdash \{\text{Prov}(\psi_i^*) \rightarrow \psi_i^* \mid i = 1, \ldots, n\} \rightarrow \phi^*\) for all realizations \(*\). Because \(\text{PA}^+\) contains all instances of \(\text{Prov}(\psi_i^*) \rightarrow \psi_i^*\), we conclude that \(\text{PA}^+ \vdash \phi^*\) for all realizations \(*\). This means that \(\text{GLS}\) is arithmetically sound with respect to \(\text{PA}^+\).

We conclude that \(\text{GLS}\) is arithmetically complete and arithmetically sound with respect to \(\text{PA}^+\). Therefore, \(\text{GLS}\) corresponds to \(\text{PA}^+\).

\textbf{(2) How does \(\text{PA}^+\) prove the consistency of \(\text{PA}\)?} Note that \(\text{PA}^+\) consists of all theorems of \(\text{PA}\) and some extra theorems. One of these extra theorems is \(\text{Prov}(\bot) \rightarrow \bot\), where \(\text{Prov}\) denotes provability in \(\text{PA}\). It follows that \(\neg \text{Prov}(\bot)\), which means that \(\text{PA}\) is consistent, is proved in \(\text{PA}^+\).

\textbf{(3) Why is \(\text{PA}^+\) the counterpart of \(T_0\) in Anderson’s system?} First we need to know what it means that \(\text{PA}^+\) is the counterpart of \(T_0\). We consider an article by Poggiolesi [2007], who explains that there are two ways to interpret the correspondence of \(\text{GLS}\) with \(\text{PA}^+\). She argues that both interpretations are incorrect because they imply \(\text{PA}^+ \not\equiv T_0\). We don’t
think that $\text{PA}^+ = T_0$ is meant by stating that “$\text{PA}^+$ is the counterpart of $T_0$”. Egré [2005, p. 26] also talks about $T'$, $U'$ and $I'$ as counterparts of $(T)$, $(U)$, and $(I)$ (described in Section 3.2), where in $T'$, $U'$ and $I'$, $K$ is replaced by $K'$ as the knowledge-plus predicate defined in Section 1.4.2. Thus $(T)$ is the axiom scheme $K(\overline{\phi}) \to \phi$ and $T'$ is $K'(\overline{\phi}) \to \phi$. It is not the case that $(T) = T'$, so we think Egré also does not mean to say that $\text{PA}^+ = T_0$. In addition to that, Egré [2005, p. 32] considers some axiom scheme which is “stronger than its tentative propositional counterpart”, from which we can also conclude that an axiom scheme which is the counterpart of another scheme is not necessarily as strong as, so not necessarily equivalent to, this other scheme. We think $\text{PA}^+$ being the counterpart of $T_0$ means that $\text{PA}^+$ and $T_0$ contain only axioms which are one another’s counterparts, like the axiom schemes $\Box\phi \to \phi$ from $\text{PA}^+$ and $K_0(\overline{\phi}) \to \phi$ from $T_0$. The counterpart axioms do not need to be equivalent or of the same strength.

Poggiolesi claims that $T_0$ contains the epistemic closure principle $[K(\overline{\phi}) \land I(\overline{\phi}, \overline{\psi})] \to K(\overline{\psi})$ while $\text{PA}^+$ does not. $\text{PA}^+$ does contain $[K(\overline{\phi}) \land (K(\overline{\phi}) \to \overline{\psi})] \to K(\overline{\psi})$, but these two schemes are “only equivalent (…) in the presence of the translation of the rule of necessitation, that is, as we already said, a rule of $\text{PA}^+$” Poggiolesi [2007, p. 161]. We agree with Poggiolesi that the schemes are not equivalent, but that does not mean that they are not counterpart of each other. The epistemic closure principle is stronger than the scheme in $\text{PA}^+$. We argued that counterpart axioms do not need to be of the same strength, so we see no reason to let this difference in the two axiom schemes lead to the idea that $\text{PA}^+$ is not the counterpart of $T_0$.

We think that $\text{PA}^+$ is the counterpart of $T_0$, because $\text{PA}^+$ extends $\text{PA}$ in an analogous way to how $T_0$ extends $Q$. To arrive at $T_0$ from $Q$, all instances of $K_0(\overline{\phi}) \to \phi$ are added for $\phi \in L_\omega$. To arrive at $\text{PA}^+$ from $\text{PA}$, all instances of $\text{Prov}(\overline{\phi}) \to \phi$ are added for $\phi$ in the language of $\text{PA}^+$, and modus ponens is applied. Like $\text{PA}$, $Q$ is closed under modus ponens. Since only instances of $K_0(\overline{\phi}) \to \phi$ are added to $Q$ to get $T_0$, and $K_0$ is not in the language of $Q$, we do not need to add anything else to $T_0$ to make sure that it is closed under modus ponens too. So because $\text{PA}^+$ is an extension of $\text{PA}$ in the same way as $T_0$ is an extension of $Q$, we conclude that $\text{PA}^+$ is the counterpart of $T_0$.

(4) What is the implicit hierarchy that is present within $T_0$ of Anderson’s system? Anderson defined $T_0 = Q \cup \{K_0(\overline{\phi}) \to \phi \mid \phi \in L_\omega\}$ and $T_{i+1} = T_i \cup \{K_{i+1}(\overline{\phi}) \to \phi \mid \phi \in L_\omega\}$ (for $i \in \omega$, $i \neq 0$). Alternatively, he could have defined $T_0 = Q$ and $T_{i+1} = T_i \cup \{K_i(\overline{\phi}) \to \phi \mid \phi \in L_\omega\}$. So $T_0$ contains an implicit hierarchy in the sense that this first part $T_0$ of the hierarchy $(T_i)_{i \in \omega}$ is already the small hierarchy of two systems $Q$ and $T_0$ itself.

We explained four claims by Egré, and it seems that these should explain
something about the similarity between Anderson’s $T_0$ and Solovay’s GLS. In Solovay’s system, we do not have $\Box(\Box \phi \rightarrow \phi)$, because the necessitation rule is not applied to instances of the reflection principle $\Box \phi \rightarrow \phi$. Do we have something like this for Anderson’s system? $T_0$ contains all instances of $K_0(\overline{\phi}) \rightarrow \phi$, just like GLS contains all instances of $\Box \phi \rightarrow \phi$ and PA$^+$ contains all instances of $\text{Prov}(\overline{\phi}) \rightarrow \phi$. Similar to the fact that we are not allowed to apply necessitation on theorems of GLS and PA$^+$ in order to get instances of counterparts of ($U$), we cannot apply necessitation within $T_0$ to get instances of $K_0(K_0(\overline{\phi}) \rightarrow \phi)$. The only kind of necessitation which can be applied in Anderson’s system to arrive at something like ($U$), is a rule which concludes $K_{n+1}(K_n(\overline{\phi}) \rightarrow \phi) \in T_{n+1}$ from $K_n(\overline{\phi}) \rightarrow \phi \in T_n$. This one does not result in an instance of ($U$) that can be used to derive the knower paradox.

Egré [2005, p. 45] calls Anderson’s hierarchy “a generalization to all the finite degrees of the separation of axiom schemata reflected in Solovay’s system”. We think that both Anderson’s and Solovay’s systems clearly indicate the rejection of the principle ($U$), implying that the knower paradox cannot be derived in these systems in the way it was originally done by Kaplan and Montague [1960]. The similarity between Anderson’s system and GLS can be argued by stating that both systems, in their own way, reject the application of the necessitation rule of inference to the reflection principle ($T$). PA$^+$ is used to show the similarity between the systems in a formal way.

Egré [2005, p. 43] adds two last sentences before his concluding remarks. “In GLS, the [necessitation rule] allows to iterate schemata [(A3)] and [(A4)] arbitrarily many times. But the reflection principle [(T)] cannot be iterated systematically, thereby preventing the appearance [of] the knower paradox.” If we consider GLS as a set of theorems for which only the inference rule modus ponens holds, this seems incorrect. However, if Egré considers GLS as a system containing the axioms of GL together with $\Box \phi \rightarrow \phi$ for which the necessitation rule only applies to the axioms of GL and modus ponens to all axioms, than the contents of the quotation is correct. In GL, necessitation can be applied to (A3) and (A4), but in GLS, there is no necessitation rule available which can be applied to principle ($T$). This is why $\Box(\Box \neg D \rightarrow \neg D)$ cannot be derived from $\Box \neg D \rightarrow \neg D$ in Solovay’s system and can thus not be used to derive the knower paradox.

### 3.4 Summary

In this chapter, we explained the three different solutions to the knower paradox described by Egré [2005]. The different solutions reject different parts of the derivation of the knower paradox by Kaplan and Montague [1960] (see Section 1.4.1, Page 14). Skyrms abandons the validity of the statement $D \leftrightarrow K(\neg D)$ as the first step of the derivation, Anderson’s solu-
tion prevents the conclusion \( D \) in Step (12), and Solovay’s solution forbids axiom scheme \( U \) such that no instance of it can be used in Step (7).

All three solutions use the notion of provability, and the goal of this thesis is to explain to what extent the knower paradox can be solved using provability logic. In the next chapter, we will discuss the quality of the three discussed solutions.
4 The Quality of Egré’s Solutions

After an introduction to the knower paradox (Chapter 1), a summary of provability logic (Chapter 2), and an explanation of some solutions to the knower paradox described by Paul Egré [2005] (Chapter 3), we discuss the quality of Egré’s solutions. We do this by arguing whether or not the solutions satisfy the requirements formulated by Susan Haack [1978]. Remember from Section 1.2.1 that she states three requirements for solutions to paradoxes. The first criterion holds that the solution should give a consistent formal theory indicating which premise(s), principle(s) of inference, of set of theorems from the theory in which the paradox is formulated should be disallowed. This is the formal part of the solution. As a second requirement, the solution should have a philosophical part indicating objections to the rejected premise(s) or principle(s) of inference, independent of its leading to the paradox. In addition, Haack’s third requirement states that a solution should not be too broad or too narrow. Egré’s solutions use the systems by Skyrms [1978], Anderson [1983] and Solovay [1976], described in respectively Sections 3.1, 3.2 and 3.3. We argue whether or not these solutions satisfy the described requirements.

4.1 The Quality of Skyrms’ Theory as a Solution

In Section 3.1, Skyrms’ theory $T_\omega = \cup_{n \in \omega} T_n$ was described. Egré considers this as a solution to the knower paradox. We discuss the extent to which this solution satisfies the requirements by Haack [1978].

4.1.1 The Formal Part of Skyrms’ Theory as a Solution

As we discussed in Section 3.1, Skyrms [1978] supposes to treat modalities in a metalinguistic way without allowing certain self-referential statements. The knower sentence, satisfying $D \leftrightarrow K(\neg D)$, is not contained in Skyrms’ system $T_\omega$, so the knower paradox cannot be derived in the original way, described in Section 1.4.1. In Section 1.4, two other versions of the knower paradox are described. The derivations of those other versions use the first formulation, so if the original derivation fails and there is no other derivation for this version of the paradox, then the other two derivations fail too.

Is Skyrms’ system a consistent formal theory which indicates a premise, inference principle, or set of theorems that should be disallowed in the theory in which the knower paradox was originally formulated? As we stated in Section 3.1, Egré [2005] shows that Skyrms’ system $T_\omega$ is a consistent formal system (if Q is consistent\(^{33}\)). Besides that, it indicates that theorems like $D \leftrightarrow \text{Prov}(\neg D)$ are not allowed in this new theory which describes

---

\(^{33}\)Like many mathematicians do, we assume that Robinson arithmetic Q is consistent.
knowledge. This means that Skyrms’ system satisfies Haack’s first criterion as a solution to the knower paradox.

4.1.2 The Philosophical Part of Skyrms’ Theory as a Solution

Does Skyrms’ theory also satisfy Haack’s second requirement? This requirement states that a solution should explain why the rejected set of theorems should be disallowed, independent of its leading to the paradox. In this case, we need arguments for disallowing statements like $D \leftrightarrow K(\neg D)$ in the theory that describes knowledge. The article by Skyrms [1978] is about modalities in general, but not specifically about knowledge. It starts with a reference to Quine [1953], who takes the view that the most natural construal of modalities is as predicates applying to names of sentences, so as metalinguistic predicates\(^{34}\). This is an argument for treating modalities metalinguistically, but not for disallowing $D \leftrightarrow K(\neg D)$.

In addition, Skyrms [1978, p. 386–387] argues that “a metalinguistic approach that avoids self-reference via a hierarchy of metalanguages leads straightforwardly to natural interpretations of S-4 and S-5”. Remember from Section 3.1 that Skyrms interpreted his $* \in *Q(S)$ as ‘is valid’ or as ‘is provable’, where we discussed only the provability interpretation. This provability interpretation leads to an interpretation of $\textbf{S4}$, which consists of the axioms and rules of inference of $\textbf{S5}$ except for axiom scheme (A5) (see Section 1.3). This means that the modal principles which hold for language $L_0$, defined as finitary language containing propositional calculus, are exactly the principles of $\textbf{S4}$. So arguments for accepting $\textbf{S4}$ as a system to describe knowledge are also arguments for accepting Skyrms’ system, but this does not directly indicate why we should disallow self-referential statements like $D \leftrightarrow K(\neg D)$. So we do not see arguments for disallowing the rejected set of theorems independent of its leading to the paradox, which means that Haack’s second requirement is not satisfied until there do rise these kind of arguments.

4.1.3 The Scope of Skyrms’ Theory as a Solution

Haack’s third requirement states that a solution to a paradox should not be too broad or too narrow. As we discussed in Section 1.2.1, a solution which is consistent satisfies the requirement that it should not be too narrow. As we have seen in Section 4.1.1, Skyrms’ system is consistent. So this system is not too narrow.

Skyrms’ system $T_\omega$ can be too broad as a solution to the knower paradox, namely if it does not contain some non-paradoxical statement which

\(^{34}\)Quine [1953] claims that the treatment of modalities as sentential operators can be reduced to the metalinguistic treatment, but Montague [1963] calls this into question. Skyrms [1978] shows that Quine is right.
is available in $L_\omega$. We cannot check this for every non-paradoxical statement, but we do this for the G"{o}del sentence $G$, with $G \iff \neg \text{Prov}('G')$. This sentence is relevant for a solution to the knower paradox, because like the knower sentence it is also a self-referential sentence about provability. Since $D$ is in $L_\omega$, and $L_\omega$ is closed under Boolean operators, it follows that $G \iff \neg \text{Prov}('G') \in L_\omega$.

Is $G \iff \neg \text{Prov}('G')$ contained in $T_\omega$? We show that it is not, which implies that Skyrms’ system is too broad as a solution to the knower paradox. The derivation is similar to the one on Page 40, where it is shown that $D \iff \text{Prov}('\neg D') \notin T_\omega$. Suppose that $G \iff \neg \text{Prov}('G') \in T_\omega = \bigcup_{n \in \omega} T_n$, then there exists some $n \in \omega$ such that $G \iff \neg \text{Prov}('G') \in T_n$. This means that for this $n \in \omega$ either $G,\neg \text{Prov}('G') \in T_n$, or $\neg G, \text{Prov}('G') \in T_n$. We consider both situations.

Suppose there exists an $n \in \omega$ such that $G,\neg \text{Prov}('G') \in T_n$. Because $G \in T_n$, $\text{Prov}('G') \in T_{n+1}$ follows from the first one of the four requirements in Theorem 3.1. Since $\neg \text{Prov}('G') \in T_n$, it follows that $\neg \text{Prov}('G') \in T_{n+1}$. Now we have both $\text{Prov}('G') \in T_{n+1}$ and $\neg \text{Prov}('G') \in T_{n+1}$, so we have a contradiction in a consistent theory. So there does not exist an $n \in \omega$ such that $G,\neg \text{Prov}('G') \in T_n$.

Now we consider the second situation. Suppose there exists an $n \in \omega$ such that $\neg G, \text{Prov}('G') \in T_n$. Because $\neg G \in T_n$, $\text{Prov}('\neg G') \in T_{n+1}$ follows by the first of the four requirements in Theorem 3.1. We also have $\text{Prov}('G') \in T_{n+1}$, because $T_{n+1}$ is an extension of $T_n$, and $\text{Prov}('G') \in T_n$. By the third requirement of Theorem 3.1, we also have $\text{Prov}('\neg G') \rightarrow \neg G \in T_{n+1}$, and $\text{Prov}('G') \rightarrow G \in T_{n+1}$, because $G, \neg G \in L_n$. Since $T_{n+1}$ is deductively closed and, $\text{Prov}('\neg G'), \text{Prov}('G') \in T_{n+1}$, it follows that $\neg G, G \in T_{n+1}$. Because $\neg G, G \in T_{n+1} \cup T_{n+1}$ is consistent, we have a contradiction. So there is no $n \in \omega$ such that $\neg G, \text{Prov}('G') \in T_n$.

So there does not exist an $n \in \omega$ such that $G, \neg \text{Prov}('G') \in T_n$ or $\neg G, \text{Prov}('G') \in T_n$. It follows that there does not exist an $n \in \omega$ such that $G \iff \neg \text{Prov}('G') \in T_n$, so we conclude $G \iff \neg \text{Prov}('G') \notin T_\omega$. This means that Skyrms’ system $T_\omega$ as a solution to the knower paradox is too broad.

We conclude that Skyrms’ system as a solution to the knower paradox does have a sufficient formal part, but the philosophical requirement by Haack is provisionally not satisfied. The third requirement, which states that the solution should not be too broad or too narrow, is partly satisfied. Skyrms’ system is too broad, but not too narrow. In Sections 4.2 and 4.3, we will consider the same criteria for the systems of Anderson and Solovay as solutions to the knower paradox.
4.2 The Quality of Anderson’s Solution

In Section 3.2, a solution to the knower paradox by Anderson [1983] is described. Like we did in Section 4.1 for Egré’s solution which used Skyrms’ system, we consider for Anderson’s solution the extent to which it satisfies Haack’s requirements.

4.2.1 The Formal Part of Anderson’s Solution

In Anderson’s hierarchy of languages, the formula $K_i(K_i(\phi) \rightarrow \phi)$ is rejected, which implies that the knower paradox cannot be derived in the way that was shown in Section 1.4.1. Does this mean the first requirement from Haack [1978] is satisfied? For this, we need a consistent formal theory indicating which premise(s), principle(s) of inference, or set of theorems from the theory in which the paradox was formulated should be disallowed. Anderson’s theory does indicate which set of theorems we should disallow, namely all instances of the axiom scheme $K(K(\phi) \rightarrow \phi)$. Is Anderson’s theory also consistent? Dean and Kurokawa [2014] write about the consistency proof Anderson sketches for his theory and state that this proof can be understood as follows.

i) We start out by interpreting $K_0(x)$ to be the set of theorems of $Z$ [for example $Z=Q$]; ii) we then take the extension of $K_1(x)$ to be the deductive closure of the extension of $K_0(x)$ plus all instances of $K_0(A) \rightarrow A$ arrived at by reflecting on the concept expressed by $K_0(x)$; iii) we then take $K_2(x)$ to be the deductive closure of the extension of $K_1(x)$ together with all instances of $K_1(A) \rightarrow A$ arrived at by reflecting on the concept expressed by $K_1(x)$; etc. [Dean and Kurokawa, 2014, p. 221].

Step ‘i)’ of the cited proof sketch can be formulated as “$V_0(K_0(\phi)) = 1$ if and only if $Q \vdash \phi$”, as we did in Section 3.2. The rest of the proof sketch is then formulated as “$V_{i+1}(K_{i+1}(\phi)) = 1$ if and only if $T_i \vdash \phi$” for $T_0 = Q \cup \{K_0(\phi) \rightarrow \phi \mid \phi \in L_\omega\}$ and $T_{i+1} = T_i \cup \{K_{i+1}(\phi) \rightarrow \phi \mid \phi \in L_\omega\}$. Step ‘i)’ implies that $T_0$ is consistent if $Q$ is consistent. This is the case, because there is no $\psi \in L_\omega$ such that $\psi \in Q$ and $\neg\psi \in \{K_0(\phi) \rightarrow \phi \mid \phi \in L_\omega\}$, or $\psi \in \{K_0(\phi) \rightarrow \phi \mid \phi \in L_\omega\}$ and $\neg\psi \in Q$. This follows because $K_0$ is not contained in the language $L_A$ of $Q$. In the same way, theories $T_i$, for $i = 1, 2, \ldots$, are consistent. So Anderson’s solution meets Haack’s first requirement.

\[\text{We still assume that Robinson arithmetic is consistent.}\]
4.2.2 The Philosophical Part of Anderson’s Solution

Haack’s second requirement concerns the philosophical part of the solution. What are the objections to the rejected scheme \((U)\), namely \(K(K(\phi) \rightarrow \phi)\)? We can divide the objections into two parts. First we consider arguments to reject the scheme \((U)\) directly, and after that we discuss a reason for disallowing the premise that there should be only one knowledge predicate. Arguments for this second rejection can be a motivation to consider to disallow scheme \((U)\). We consider arguments from articles by Anderson [1983], Poggiolesi [2007], and Dean and Kurokawa [2014]. We state four arguments for rejecting \((U)\) directly, and one for disallowing the idea of the existence of only one knowledge predicate.

(1) Interpreting Knowledge as Provability The first argument to disallow axiom scheme \((U)\) is given both by Anderson [1983, p. 350] and Poggiolesi [2007, p. 152]. This scheme is not valid in a system where provability is considered instead of knowledge. Remember that the knower paradox followed from the combination of the schemes \(K(\phi) \rightarrow \phi, K(K(\phi) \rightarrow \phi), \) and \([I(\phi, \psi) \land K(\phi)] \rightarrow K(\psi)\) \((U),\) and \((I)\) respectively. Presuming a provability interpretation, the schemes \(\text{Prov}(\phi) \rightarrow \phi\) and \([I(\phi, \psi) \land \text{Prov}(\phi)] \rightarrow \text{Prov}(\psi)\) are valid, while \((U)\), interpreted as \(\text{Prov}(\text{Prov}(\phi) \rightarrow \phi)\), is not. So interpreting knowledge as provability implies that \((U)\) should be disallowed.

(2) Do not Reject \((I)\) and \((T)\) The second argument is only mentioned by Poggiolesi [2007, p. 151–152], and is not independent of solving the paradox. It argues that the other two schemes, \((I)\) and \((T)\), are more difficult to doubt. As stated in Section 1.4.2 and 1.4.3, Cross [2001] is able to construct the knower paradox without the epistemic closure principle \((I)\), \([I(\phi, \psi) \land K(\phi)] \rightarrow K(\psi)\), so disallowing this axiom scheme does not solve the paradox.

In addition, Poggiolesi argues that the reflection principle \((T)\), \(K(\phi) \rightarrow \phi\), is the “most natural and intuitive” scheme, which is part of the definition of knowledge as justified true belief. Poggiolesi refers to Campbell [1883], stating that this standard conception of knowledge was first introduced by Plato in the Theaetetus. She also states that “it is implausible to reject the reflection principle without rejecting knowledge itself” [Poggiolesi, 2007, p. 152]. So the only axiom scheme which we can reject from \((T)\), \((U)\), and \((I)\), is principle \((U)\).

(3) \((U)\) Affirms Knowledge As a third reason, Poggiolesi [2007, p. 152–153] explains that in contrast with the other two principles, \((U)\) affirms something about knowledge. Both \((T)\) and \((I)\) are implications, condition-
ally stating something about knowledge. These schemes can hold in a system in which nothing is known at all. However, \((U)\) just states the knowledge of something unconditionally. The set of known statements cannot be empty in a system in which \((U)\) holds, while this is possible in a system where only \((T)\) and \((I)\) hold.

(4) \((U)\) Represents Epistemic Non-Conservativity A last argument for rejecting the scheme \((U)\) directly is given by Dean and Kurokawa [2014, p. 217]. They explain that \((U)\) represents epistemic non-conservativity. Some system \(S\) is epistemically conservative over system \(Z\) if \(S\) does not derive \(K(\phi)\) unless \(\phi\) is derivable in \(Z\), so if \(S \vdash K(\phi)\) implies \(Z \vdash \phi\) [Dean and Kurokawa, 2014, p. 215]. A system is epistemically non-conservative over \(Z\) if it is not epistemically conservative over \(Z\).

The principle \(K(K(\phi) \rightarrow \phi)\) contains instances of epistemic non-conservativity, for example if we consider \(\phi\) to be \(\text{Prov}_Z(\bot) \rightarrow \bot\), which represents the consistency of \(Z\) and is abbreviated by \(\text{Con}(Z)\). If we consider \(Z\) to be PA, then by Gödel’s second incompleteness theorem, \(Z \vdash \text{Con}(Z)\) cannot be true. Suppose we take \(S\) as \(Z\) together with \((T)\), \((U)\), \((I)\) and \(K(B)\) for some sentence \(B\). Then we can show that \(S \vdash K(\text{Con}(Z))\) is true, which implies that \(S\) is epistemically non-conservative over \(Z\). Why does \(S \vdash K(\text{Con}(Z))\) hold? We interpret \(K(x)\) as \(\text{Prov}_Z(x)\), so \((U)\) is interpreted as \(\text{Prov}_Z(\text{Prov}_Z(x) \rightarrow x)\) and \(K(\text{Con}(Z))\) as \(\text{Prov}_Z(\text{Con}(Z))\), so as \(\text{Prov}_Z(\text{Prov}_Z(\bot) \rightarrow \bot)\). Since under this interpretation the last statement is an instance of \((U)\), we can conclude for this interpretation that \(S \vdash K(\text{Con}(Z))\) holds.

So both \(Z \not\vdash \text{Con}(Z)\) and \(S \vdash K(\text{Con}(Z))\) hold, which means that \(S\) is epistemically non-conservative over \(Z\). Proposition 5.2 of [Dean and Kurokawa, 2014, p. 215] states that \(S_0\), which is \(S\) without \((U)\), is epistemically conservative over \(Z\). Therefore, rejecting \((U)\) leads from epistemic non-conservativity to epistemic conservativity.

(5) Anderson’s ‘Intuitive Motivation’ After four arguments for directly rejecting scheme \((U)\), we give an argument for accepting more than one knowledge predicate. According to Egré [2005, p. 38], Anderson’s philosophical motivation to reject axiom scheme \((U)\) is “the idea that the two occurrences of the knowledge predicate (…) need not be on the same level; what is known at one stage is likely to modify what can be known at the next stage”. This can be seen in the “intuitive motivation” for a hierarchy of knowledge predicates by Anderson [1983, p. 348–349], which we discussed in Section 3.2. Remember that this contained the system \(Q'\) based on \(Q\) and the set of statements \(K_0\) known by some person, whom we call Oscar. There are provable statements which cannot be in \(Q'\), for example the Gödel sentence \(G\). After reflecting on some arguments, Oscar can un-
nderstand the proofs of these statements that cannot be in $Q'$, the system representing Oscar’s knowledge. Therefore, it was concluded that it seems intuitive to accept different knowledge levels representing different parts of Oscar’s knowledge. This idea of different levels of knowledge is an argument to reject axiom scheme $K(K(\bar{\phi}) \rightarrow \phi)$ and accept $K_{i+1}(K_i(\bar{\phi}) \rightarrow \phi)$.

Anderson could have paid more attention to some details of his argument. At first, we consider the fact that he uses two systems which are defined as $Q'$. The first one is “an extension of Robinson arithmetic containing additional non-logical constants”. Then Anderson [1983, p. 348] constructs another formal system “whose theorems are the sentences derivable from the axioms of Robinson arithmetic together with the sentences of $K_0$”, where $K_0$ is defined as “the things expressible in [the first defined version of] $Q'$ which are known by a certain person”. We are wondering whether Anderson decides to “call this axiomatic system $Q'$ also”. Secondly, we wonder what Anderson [1983, p. 348] means with his definition of $K_0$. It seems that “things expressible in $Q'$ which are known by a certain person” can be false statements, but later in his argument he uses that all sentences of $K_0$ are true (“things which are known to Oscar, and hence are true”). We assume that, in this ‘intuitive motivation’ for accepting more than one knowledge predicate, Anderson uses the premise that knowledge implies truth, such that “things expressible in $Q'$ and known by a certain person” cannot be false statements.

Anderson [1983, p. 349] admits that his intuitive argument is “rather suspicious, but [he] can’t see what’s wrong with it”. Poggiolesi [2007] argues that there is indeed something wrong with this argument. Her main reason is that Anderson uses two different notions of proof. There is the syntactic proof within a certain system, which is “a finite sequence of sentences, each of which either is an axiom of the system, or is immediately deducible from the earlier ones by one of the rules of inference” [Poggiolesi, 2007, p. 155]. Then there is the notion of an absolute proof, which uses reasoning which cannot be formalized in the system. The reasoning from which we conclude that Gödel sentence $G$ is true is an example of an absolute proof, while the fact that $G$ is not in $K_0$ follows by a syntactic proof in $T_0$.

Poggiolesi [2007, p. 157] states that “there is no reason for changing the notion of proof on which [Oscar’s] knowledge is based” and concludes that “Anderson cannot give a definite and clear notion of knowledge”.

We do not agree with Poggiolesi that the use of different notions of proof is an important problem for Anderson’s solution. On the contrary, the idea that knowledge can be acquired in different ways supports the philosophical part of Anderson’s solution. Since we can gain knowledge via both syntactical proofs and absolute ones, it is plausible to define at

\[\text{Starting with } T_0 \vdash G \leftrightarrow \neg K_0(\overline{G}), \text{ one can use } T_0 \vdash K_0(\overline{G}) \rightarrow G \text{ and } T_0 \vdash K_0(\overline{G}) \vee \neg K_0(\overline{G}) \text{ to deduce } T_0 \vdash G \text{ and thus } T_0 \vdash \neg K_0(\overline{G}).\]
least two different kinds, or levels, of knowledge. Why do we need more than two levels? Following the reasoning for the existence of knowledge $K_1$ of level 1, we can construct system $Q''$ as $Q'$ together with all sentences from $K_1$. In this system, we can conclude that statement $G'' \notin K_1$, where $G'' \leftrightarrow \neg \text{Prov}(G')$ for $\text{Prov}$ denoting provability in $Q''$, but $G' \in K_2$, where $K_2$ is the set of sentences known at level 2. Like the proof that $G \notin K_0$, the proof that $G'' \notin K_1$ is not syntactical but absolute. In the same way, the existence of even higher knowledge levels can be argued. Except for level 0, the arguments for the knowledge levels are all based on absolute proofs, but if we are allowed to apply the reasoning for the existence of $K_1$, we are allowed to use the similar reasoning for the existence of the higher levels. Only, we would expect again some different kind of proof, other than syntactical and absolute. Since the argumentation for every additional knowledge level, except for the first one (level 1), needs a kind of proof which was already used in the argumentation for some lower knowledge level, we do not consider Anderson’s ‘intuitive motivation’ as a sufficient argument for the acceptance of an infinite number of knowledge predicates.

In their article *The Paradox of the Knower revisited*, Dean and Kurokawa [2014, p. 199] describe to what extent they agree with Anderson’s ‘intuitive motivation’ for a hierarchy of knowledge levels. They argue that “while the knower paradox can be understood to motivate a distinction between levels of knowledge, it does not provide a rationale for recognizing a uniform hierarchy of knowledge predicates in the manner suggested by Anderson”. They explain that provability implies knowledge and they support a provability interpretation of knowledge. The existence of different axiomatic proof systems like $Q$ and $\text{PA}$ might lead to different levels of knowledge. However, a problem with Anderson’s theory is the infinite number of knowledge predicates. Dean and Kurokawa [2014, p. 204] wonder whether it would be possible to find a correspondence between different knowledge levels and categories of knowledge traditionally recognized in general epistemology, like logical knowledge, a priori knowledge, and a posteriori knowledge. They state that “nothing in our everyday practices leads us to recognize the existence of infinitely many uniformly parametrized notions of knowledge”. In addition, they suggest that knowledge level $K_1$ already contains a lot of statements which are not contained in $K_0$, for example all theorems of PA are in $K_1$ if $Q'$ is considered as $I\Delta_0 + EXP$ [Dean and Kurokawa, 2014, Theorem 6.2]. This implies that we are “uncertain whether $K_0(x)$ itself in fact describes any pre-theoretically recognizable concept of knowledge”, and “uncertainty can only grow when we attempt to assign interpretations to $K_1(x), K_2(x), \ldots$” [Dean and Kurokawa, 2014, p. 223]. So it seems appro-

---

Dean and Kurokawa [2014, p. 204] call this system EA. It is elementary arithmetic containing $Q$, induction for formulae which contain only bounded quantifiers, and the assertion that the exponentiation function is total.
appropriate to accept more than one level of knowledge, but not via Anderson’s system.

Together, the five arguments that we discussed, some given by Anderson, some by Poggiolesi, and some by Dean and Kurokawa, can form the philosophical part of Anderson’s solution to the knower paradox. This philosophical part indicates objections to the rejected principle \((U)\). Not all of the arguments can be used to satisfy Haack’s second criterion, for example, because the reason to reject the principle \((U)\) should not depend on the existence of the paradox. This is why the second argument, which explains that \((T)\) and \((I)\) are more difficult to reject, fails.

Our discussion of the last argument, about the ‘intuitive motivation’ to accept more than one knowledge predicate, ended in the conclusion that more than one knowledge predicate should be accepted, but not in the way it was done by Anderson. Therefore, this argument fails as an argument for Anderson’s system as a solution to the knower paradox. We are also not convinced by the third argument, which was Poggiolesi’s explanation that we should disallow \((U)\) because, in contrast to the other axiom schemes, it affirms something about knowledge. It is true that this distinguishes \((U)\) from the other schemes, but why would we disallow this, independent of the existence of the paradox?

Besides these arguments, there are two arguments left which could be accepted to satisfy Haack’s second requirement. The first reason to reject \((U)\) was because it is not valid if we interpret knowledge as provability, while the fourth one stated that rejecting \((U)\) leads from epistemic non-conservativity to epistemic conservativity. We think these two arguments together form enough reason to disallow \((U)\), independent of the existence of the knower paradox. Therefore, we conclude that Anderson’s system satisfies Haack’s second criterion.

### 4.2.3 The Scope of Anderson’s Solution

Haack’s third requirement states that a solution to a paradox should not be too broad or too narrow, which means that it should not contain any paradoxes, but it has to contain all non-paradoxical statements which can be formulated in the languages of the regarded system. Like we explained in Section 1.2.1, a given system is not too narrow if it is consistent. This is the case for Anderson’s solution to the knower paradox, so Anderson’s solution is not too narrow.

A statement which is potentially able to show that a solution to a paradox is too broad is Gödel sentence \(G\), which satisfies \(G \leftrightarrow \neg K_i(G)\) for some \(i \in \omega\). Suppose that \(T_{i=j}\) is the first theory of Anderson’s hierarchy in which \(G\) occurs. Then \(T_{j-1} \not \vdash G\), so \(V_j(K_j(G)) = 0\). Therefore, we have \(V_j(G) = 1\) and \(V_j(\neg K_j(G)) = 1\), so \(V_j(G \leftrightarrow \neg K_j(G)) = 1\). This means that
there is indeed some \( i \in \omega \), namely \( i = j \), for which \( G \leftrightarrow \neg K_i(G) \in T_i \). Our provisional conclusion is that Anderson’s system satisfies Haack’s third requirement, but still someone might find out at some stage that it does not.

Summarizing, Anderson’s system satisfies both the formal and the philosophical requirements formulated by Haack. The system is not too narrow and provisionally not too broad, so the third requirement is provisionally met. So Anderson’s system, together with arguments by Poggiolesi [2007] and Dean and Kurokawa [2014], satisfies all of Haack’s requirements on solutions to paradoxes at least provisionally.

The requirement on the philosophical part of the solution is satisfied by Anderson’s system, using one additional argument for rejecting (\( U \)) by Dean and Kurokawa [2014]. However, Dean and Kurokawa’s conclusion is that Anderson’s system is not sufficient for solving the knower paradox, while our conclusion is that it does meet all of Haack’s requirements at least provisionally. Does this mean that Haack’s criteria are not sufficient for finding out the quality of a solution to a paradox? Or should we have added that the premise ‘there is exactly one knowledge predicate’ is disallowed, such that we would need to find philosophical arguments for rejecting this in order to meet the second requirement? This last option would not solve the problem, because Dean and Kurokawa also have arguments for rejecting this premise. The aspect that distinguishes their idea of a solution to the knower paradox from the solution by Anderson, is the number of knowledge predicates which replace the single knowledge predicate in the system of Kaplan and Montague. So maybe a requirement should be added which requires philosophical reasons to accept premises which replace a rejected premise.

4.3 The Quality of Solovay’s Theory as a Solution

In Chapter 3, three different systems which Egré [2005] considers as solutions to the knower paradox were discussed. The quality of the first two systems, respectively by Skyrms and Anderson, was discussed in Sections 4.1 and 4.2. In the current section, we check whether the third system, by Solovay, satisfies the three requirements on solutions to paradoxes by Haack [1978].

4.3.1 The Formal Part of Solovay’s Theory as a Solution

First of all, the solution should contain a consistent formal system indicating an unacceptable premise, principle of inference, or set of theorems. Solovay’s formal system GLS indicates the rejection of \( K(K(\phi) \rightarrow \phi) \), which is achieved by disallowing the necessitation rule to apply to the reflection principle \( K(\phi) \rightarrow \phi \). Is GLS consistent? Solovay [1976] proved that GLS
is arithmetically sound with respect to the standard model. Since truth is a model implies consistency, GLS is consistent. So Haack’s first requirement on solutions to paradoxes is satisfied.

4.3.2 The Philosophical Part of Solovay’s Theory as a Solution

To satisfy Haack’s second requirement, there needs to be an argumentation for rejecting $K(K(\phi) \rightarrow \phi)$ or for disallowing the necessitation rule to apply to the reflection principle $K(\phi) \rightarrow \phi$. This argumentation should be independent of the existence of the knower paradox. Since Solovay [1976] did not consider GLS within the context of the knower paradox, a potential argumentation by him will not be dependent on the knower paradox. However, Solovay’s article is about provability, but not about knowledge, so we do not find arguments for rejecting $K(K(\phi) \rightarrow \phi)$ there. Considering provability, there are reasons to reject $\text{Prov}(\text{Prov}(\phi) \rightarrow \phi)$. Remember from Section 2.4.3 that Löb’s theorem states that $\text{PA} \vdash \text{Prov}(\text{Prov}(\phi) \rightarrow \phi) \rightarrow \text{Prov}(\phi)$. This implies that if $\text{Prov}(\text{Prov}(\phi) \rightarrow \phi)$ is accepted as an axiom scheme, then $\text{Prov}(\phi)$ holds for every statement $\phi$, even for false statements. This is an argument to accept GLS as a system to interpret provability, but not directly to accept it as a system to interpret knowledge.

Egré [2005, p. 42] argues that GL can be seen as a “system formalizing the knowledge of an ideal mathematician recursively generating all the theorems of PA and reflecting on the scope of his knowledge”. If we want to keep axiom ($T$), $K(\phi) \rightarrow \phi$, in our representation of knowledge, we should make sure that the necessitation rule is not allowed to apply to ($T$) in order to prevent the knower paradox. This results in the system GLS. The only reason we can find in [Egré, 2005] for accepting exactly this system is not independent of the existence of the paradox, because we disallow the necessitation rule to apply to ($T$) just to prevent the paradox. Therefore, Haack’s second requirement is not satisfied for Solovay’s system. Still reasons to let a knowledge predicate satisfy the axioms of GLS can be found. Finding such reasons would imply that the second criterion is satisfied.

4.3.3 The Scope of Solovay’s Theory as a Solution

Haack’s third requirement states that a solution to a paradox should not be too broad or too narrow. Like we did in the evaluations of both Skyrms’ and Anderson’s system (see Sections 4.1 and 4.2), we conclude that a system is not too narrow if it is consistent. Solovay’s system is consistent, so it is not too narrow and we conclude that this part of Haack’s third requirement is satisfied.

We conclude provisionally that a solution is not too broad if we do not find an example of a theorem which should be, but is not, a theorem of the system. We cannot check this for all theorems, but we consider the
same example as in Sections 4.1 and 4.2. Gödel sentence \( G \) in PA satisfies \( \text{PA} \vdash G \leftrightarrow \neg \text{Prov}(G) \). Is there a sentence \( G \) in GLS that satisfies \( G \leftrightarrow \neg \Box G \)? Yes there is, namely \( \neg \Box \bot \). This formula \( \neg \Box \bot \) is in GLS, because GLS is consistent. The formula \( \neg \Box \bot \leftrightarrow \neg \Box \neg \Box \bot \) is in GL, and thus in GLS. The formula \( \neg \Box \neg \Box \bot \rightarrow \neg \bot \) is in GL, because GL \( \vdash \Box \bot \rightarrow \Box \neg \Box \bot \). Since \( \Box \neg \Box \bot \rightarrow \Box \bot \) is an axiom of GL, it follows that GL \( \vdash \Box (\neg \Box \bot ) \rightarrow \Box \bot \). Therefore, also \( \neg \Box \bot \rightarrow \neg \Box \neg \Box \bot \) is contained in GL. So there is some \( G \), namely \( \neg \Box \bot \), which satisfies GLS \( \vdash G \leftrightarrow \neg \Box G \), which means that the Gödel sentence is a theorem of Solovay’s system. So provisionally, Solovay’s system is not too broad.

Summarizing the discussion about the quality of Solovay’s system as a solution to the knower paradox, Haack’s first requirement is satisfied and the solution falls provisionally short of the second criterion. The third criterion is provisionally met, because the solution is not too narrow and provisionally not too broad.

4.4 Summary

We discussed the quality of the theories of Skyrms [1978], Anderson [1983] and Solovay [1976], which were described in Chapter 3. Consider Table 1 for a summary of this discussion.

<table>
<thead>
<tr>
<th></th>
<th>Skyrms</th>
<th>Anderson</th>
<th>Solovay</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Formal</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2. Philosophical</td>
<td>x...</td>
<td>...</td>
<td>x...</td>
</tr>
<tr>
<td>3. Scope</td>
<td>Not too narrow</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Not too broad</td>
<td>x</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 1: Summary of Chapter 4. The symbol ‘✓’ means ‘satisfied’, ‘x’ means ‘not satisfied’, and the addition of ‘...’ means ‘provisionally’.

The systems of Skyrms, Anderson, and Solovay all satisfy Haack’s first requirement. The second requirement is met by Anderson’s system in combination with arguments by Poggiolesi [2007] and Dean and Kurokawa [2014], but provisionally not by Skyrms’ and Solovay’s system. We denote that Anderson’s system only provisionally meets this requirement, because there could always arise arguments which take the edge off the current arguments.

Because all systems we considered are consistent, all solutions are not too narrow. Finally, we tried to find out whether the solutions are too broad. To do this, we considered the Gödel sentence \( G \), satisfying \( G \leftrightarrow \neg K(\overline{G}) \). We concluded that \( G \) is not in Skyrms’ system, but it is in Anderson’s and Solovay’s systems. So Skyrms’ solution is too broad, but provisionally, the other two solutions are not. So far, the best solution is Anderson’s system, which best meets Haack’s requirements.
5 Discussion and Conclusion

In this thesis, we want to answer the following question. To what extent can provability logic be used to solve the knower paradox? The knower paradox and provability logic were described in Chapter 1 and Chapter 2 respectively. We considered three possible provability interpretations in Chapter 3, and the quality of these interpretations as solutions to the knower paradox was discussed in Chapter 4. Our provisional conclusion was that the interpretation of Anderson [1983] is the best of these three, because it is the one which best meets the requirements on solutions to paradoxes by Haack [1978]. The two other solutions by Égré [2005] we discussed, which use systems by Skyrms [1978] and Solovay [1976], could be improved. In this final chapter we consider to what extent this is possible, and we comment on the idea of interpreting knowledge as provability in general.

5.1 Improving Égré’s Solutions?

In Chapter 4, three systems that represent provability were used by Égré [2005] to interpret knowledge. We discussed to what extent these solutions satisfied the requirements by Haack, described in Section 1.2.1. One of the solutions, namely the one by Anderson [1983], satisfies all these requirements at least provisionally. The other two solutions, by Skyrms [1978] and Solovay [1976], do not satisfy the requirement on the philosophical part of the solution. In this section, we discuss whether we can improve this part of each of these two solutions.

Improving the Philosophical Part of Skyrms’ System as Solution

In Section 4.1, we considered whether Skyrms’ system $T_\omega$ as solution to the knower paradox satisfies Haack’s requirements. It turned out that the philosophical part is provisionally not sufficient and that it is too broad. To improve Skyrms’ provability interpretation as solution to the knower paradox, we need to find philosophical arguments for disallowing the statement $D \leftrightarrow K(\neg D)$ to be a theorem in a system which represents knowledge.

Looking for a reason to disallow an agent to state that she knows the negation of the sentence she states, we cannot think of any other reason than just because it leads to a paradox. This is not a reason which can be used in the philosophical part of the solution. It is not strange that we cannot find a reason which completes this part of the solution, because in Section 1.2.1, we concluded that a solution which is too broad, has no sufficient philosophical part. Therefore, we cannot improve Skyrms’ system as a solution to the knower paradox.

Improving the Philosophical Part of Solovay’s System as Solution

In Section 4.3, we discussed the quality of using Solovay’s system as solution
to the knower paradox. It appeared to be the case that Solovay’s system is consistent and indicates that the application of necessitation to the reflection principle \((T)\) should be rejected. Therefore, it satisfies Haack’s first requirement. The second requirement however, for which arguments for disallowing the rejected premise, principle of inference, or set of theorems are needed, was not satisfied yet. To improve Solovay’s system as solution to the knower paradox, we need to find philosophical arguments for disallowing the application of necessitation to the reflection principle \((T)\). If this application of necessitation is not allowed, we do not have axiom scheme \((U)\), so the knower paradox cannot be derived as we did in Section 1.4.1 on Page 14.

As we have explained in the evaluation of Anderson’s system, there are enough reasons to disallow axiom scheme \((U)\) (see Section 4.2.2). The two arguments which we accepted to satisfy Haack’s second requirement can also be used to complete Egré’s idea to use Solovay’s system as solution to the knower paradox.

The first argument we mentioned was that scheme \((U)\) is not valid in a system where provability is considered instead of knowledge. Accepting provability as interpretation of knowledge is a good reason to accept Solovay’s system as a solution to the knower paradox, since Solovay’s \(GLS\) is a system about provability which is arithmetically complete and arithmetically sound with respect to the standard model \(\omega\). This indicates that \(GLS\) describes mathematical knowledge, namely facts about provability in Peano arithmetic which are known by mathematicians.

The second argument we accepted to satisfy Haack’s requirement on the philosophical part of Anderson’s solution to the knower paradox is that rejecting \((U)\) leads from epistemic non-conservativity to epistemic conservativity. Is \(GLS\) epistemically conservative over PA? This is the case if \(GLS \vdash \Box \phi \Rightarrow PA \vdash \phi^*\) for every realization \(*\), where \(\Box\) is interpreted as knowledge. Since \(GLS\) is arithmetically sound with respect to the standard model \(\langle \omega; +, \cdot \rangle\), \(GLS \vdash \Box \phi \Rightarrow \omega \models Prov_{PA}(\sigma^*)\) for all realizations \(*\). This means that there exists a proof of \(\phi^*\) in PA for all realizations \(*\), so \(PA \vdash \phi^*\) holds for every realization \(*\). So for \(\Box\) interpreted as knowledge, \(GLS\) epistemically conservative over PA, which is an argument to accept this theory as a solution to the knower paradox.

These two arguments form a satisfying philosophical part of Egré’s idea to use Solovay’s system \(GLS\) as formal solution to the knower paradox. Therefore, we now conclude that this Solovay’s system, together with these arguments, satisfies all of Haack’s requirements at least provisionally. We explain why we prefer Solovay’s system to the one by Anderson.
Comparing the Satisfactory Solutions  Our provisional conclusion of Chapter 4 was that the interpretation of Anderson [1983] is the best of these three, because it is the one which best meets the requirements on solutions to paradoxes by Haack [1978]. We have found arguments which satisfy the philosophical part of Solovay’s system as solution to the knower paradox, so Solovay’s system satisfies all of Haack’s requirements at least provisionally, just like Anderson’s solution.

We prefer Solovay’s system to Anderson’s, because of the number of different knowledge levels. As we stated in Section 4.2.2, we did not agree with the idea of more than two knowledge levels in the way it is defined by Anderson. If we define knowledge in the way provability is defined in Solovay’s system, we have only one knowledge level.

We did agree with Anderson’s intuitive motivation to have two different knowledge levels. Do we want to have one extra knowledge level in Solovay’s system? If we indeed want this, we could add an arithmetical predicate Prov′, interpreted as provability outside PA. We would need to define this Prov′ in a way such that the new system is arithmetically complete and arithmetically sound with respect to some arithmetical model. Such bimodal logics are discussed for example by Beklemishev [1994] and Smoryński [1985, Chapter 4].

Dean and Kurokawa [2014] consider the search for even more provability predicates, which represent provability in many different axiomatic systems like Q, IΔ0 + EXP, and extensions of PA. Each different provability predicate could be used as an interpretation of all different kinds of knowledge, like logical knowledge, a priori knowledge, and a posteriori knowledge. Dean and Kurokawa express their doubts as to whether such a precise classification is possible. We agree with them, but we would like to add that it might be less doubtful whether such a classification is possible if we do not consider kinds of knowledge like ‘a priori knowledge’ and ‘a posteriori knowledge’, but ‘knowledge of statements in X’ for axiomatic systems X. In that case, we could interpret knowledge of statements in Q as ProvQ, knowledge of statements in IΔ0 + EXP as ProvIΔ0+EXP, etcetera. Whether such an interpretation of different kinds of knowledge as different kinds of provability is possible, would be an interesting question for further research.

In this section, we concluded that Skyrms’ system as a solution to the knower paradox cannot be improved in order to make it satisfy all of Haack’s requirements. To Solovay’s system, we did add some arguments that made the requirement on the philosophical part of the solution satisfactory. So now both Solovay’s and Anderson’s system satisfy all of Haack’s requirements. We argued that we prefer Solovay’s solution to Anderson’s, because we did not agree with Anderson’s motivation for more than two different knowledge levels. We now consider whether the idea that knowledge can be interpreted as provability, which is used in the philosophical part of both Anderson’s
solution and the solution which uses Solovay’s system, is arguable.

5.2 Interpreting Knowledge as Provability

Three interpretations of provability logic were discussed as solutions to the knower paradox. The three systems we considered are all used by Egré [2005] to interpret knowledge, applying a certain definition of provability. Each of the three solutions contains provability in a theory which extends Robinson arithmetic. In Skyrms’ system, $\text{Prov}(\phi)$ means ‘$\phi$ is provable in $T_\omega$’. In Anderson’s system, $K_i(\phi)$ means ‘$\phi$ is known at level $i$', which is the case for $i = 0$ if $\phi$ is provable in $Q$. In Solovay’s system, $\square \phi$ means ‘$\phi$ is provable in some theory of arithmetic, for example Peano arithmetic’. Can one maintain that the concepts of knowledge and provability coincide?

In this section, we consider some arguments for and against the idea that knowledge and provability coincide, where we mean specific kinds of knowledge and provability. We consider mathematical knowledge, namely facts about (Peano) arithmetic which are known by at least one mathematician. We say that a statement is provable if there exists a proof of it in Peano arithmetic.

First we consider why it seems plausible to interpret knowledge as provability. If some statement is provable, then there exists a proof of that statement. This proof is understood at least by the person who came up with it, so this person knows the proved statement. Thus, provability seems to imply knowledge.

According to the intuition of many mathematicians, the other way around also holds. Some statement can only be mathematical knowledge if it is also provable in PA. If some statement about (Peano) arithmetic is not provable, then there is no proof of it, so no mathematician can know the statement.

In addition, an argument for interpreting knowledge as provability is that it solves the knower paradox. Not all three discussed solutions satisfy all of Haack’s criteria, but most are satisfied at least provisionally.

Can one maintain that the concepts of mathematical knowledge and provability in Peano arithmetic coincide? We have seen that intuitively, this seems to be the case. However, there are also arguments against interpreting knowledge as provability.

---

38Only the second system, by Anderson, was used to interpret knowledge in the article in which it was published.  
39We could also require that the fact about arithmetic is known by all mathematicians instead of at least one. If we would go for ‘all mathematicians’, we necessarily need to check all mathematicians to found out that some statement is known. We prefer to require knowledge by ‘at least one mathematician’, because this implies that as soon as we find one mathematicians who knows the statement, we can conclude that the statement is known.

64
Issues that cast doubt on the intuition about the similarity between knowledge and provability are the following. We stated that the existence of a proof implies that there is a person who came up with it, but this only holds from a constructivistic point of view. According to a constructivist, proofs are constructed by mathematicians, so there exists a proof of a certain statement only if there is (or has been) some mathematician who proved it. In this way, a statement can only be provable if it is known. This also means that a statement which will be proved next year, but is not proved at the moment, is not provable yet. Considering provability in this time-dependent way seems counterintuitive, so we consider another view on the metaphysics of mathematics, namely Platonism.

According to a Platonist, a proof exists independently of mathematicians. This means that even a theorem which will be proved only next year, is provable independent of the current time. It seems to be plausible to define provability independent of time and independent of mathematicians, but knowledge does depend on time, or at least on (the existence of) mathematicians. So an argument Platonists can use against interpreting knowledge as provability is that knowledge seems to be dependent on mathematicians and on time, while provability does not.

Independent of our ideas about metaphysics of mathematics, there are statements which are known but not provable in PA. There are also statements which are provable in PA but not known, specifically if we accept the Platonistic view. An example of the first kind is the Gödel sentence \( G \) for PA, with \( G \leftrightarrow \neg \text{Prov}(\overline{G}) \). This sentence about arithmetic is not provable in PA, but via reasoning outside PA, mathematicians can gain the knowledge that \( G \) holds. The same holds for the strengthened finite Ramsey theorem\(^{40}\), whose truth can be shown in second-order arithmetic, but of which the Paris-Harrington theorem states that it is not provable in PA [Paris and Harrington, 1977].

An example of the second kind can be found by considering a theorem which has been a conjecture for some time and finally is proved in PA. We consider Catalan’s conjecture, which states that the unique solution\(^{41}\) to \( x^m - y^n = 1 \) is \( x = \pm 3 \) and \( y = 2 \). While the conjecture was stated in 1844, a full proof was first given by Mihăilescu in 2002 [Daems, 2004]. This proof is partly based on logarithmic forms and electronic computations, but Bilu [2005] shows that Catalan’s conjecture can be proved without these. Since this proof is mainly based on basic theorems about cyclotomic fields, which are provable in PA, we assume that the conjecture is provable in PA.

\(^{40}\)The strengthened finite Ramsey theorem states that for any positive integers \( n, k, \) and \( m \) one can find an integer \( N \) such that the following holds. If each of the \( n \)-element subsets of \( S = \{1, 2, 3, \ldots, N\} \) is colored with one of the \( k \) colors, then there exists a subset \( T \) of \( S \), consisting of at least \( m \) elements, such that all \( n \)-element subsets of \( T \) have the same color, and the number of elements of \( T \) is at least the smallest element of \( T \).

\(^{41}\)Assuming \( m, n \) are integers greater than 1 and \( x, y \) are both unequal to 0.
means we have an example of something which is provable in PA, but was not known before 2002. For a Platonist, the proof always existed, so the conjecture has always been provable. Before 2002, the provability of this conjecture did not imply that its contents was mathematical knowledge.

Another example is Löb’s theorem, which we discussed in Section 2.4.3. The formalized version of this theorem, PA ⊢ Prov(Prov(φ) → φ) → Prov(φ), is a statement which is provable in PA, but one which was not known for a long time. The theorem is even “utterly astonishing”, as explained by Boolos [1995, p. 54], because the mathematical gap between truth and provability is difficult to understand. Before Löb proved his theorem, it was not known that it held, but in the Platonistic view of the existence of mathematical objects like proofs, it has always been provable. So this is a second example of a theorem which was not known at a certain time, but which has been provable in PA all along.

In addition to these examples, we consider how the correspondence between knowledge and provability is discussed in the literature. This is done in the context of justification logic. A justification logic is an epistemic logic where a statement is known only if there is given some justification [Artemov and Fitting, 2012]. Instead of implicit knowledge □φ, meaning ‘it is known that φ’, we have explicit knowledge t : φ, meaning that ‘φ is justified by t’. A logic which combines implicit and explicit knowledge is S4LP by Artemov and Nogina [2005], which contains the axiom scheme (t : φ) → □φ. Since a proof is a form of justification, this justification logic is a way to connect knowledge and provability. Some criticism on this logic is the following. When trying to indicate why ‘it is known that φ’ is not true, an adherent of justification logic tries to find a mistake in the justification t for φ, while in non-justification epistemic logic, we try to find a counterexample directly against ‘it is known that φ’. If a mistake is found in justification logic, this does not necessarily imply that ‘it is known that φ’ is not true. As we see, there are differences between knowledge which we usually consider, namely implicit knowledge, and explicit knowledge in the sense of justification or provability.

So, do we conclude that, in spite of the described philosophical solutions to the knower paradox, knowledge should not be interpreted as provability? We saw the following arguments against the correspondence between knowledge and provability. The Gödel sentence for PA and the strengthened finite Ramsey theorem show that mathematicians know more than only theorems in PA. If we accept the Platonistic view of metaphysics of mathematics, then provability in PA does not depend on mathematicians and time, while mathematical knowledge does. This means that knowledge of theorems in PA only exists after some mathematician proved them, while these theorems have been provable all along. In addition, the way in which
we consider knowledge is like the explicit knowledge from justification logic, while provability is a form of implicit knowledge.

Are these reasons to conclude that knowledge should not be interpreted as provability, and thus that provability logic does not solve the knower paradox? Let us consider for a moment a system which is often used to model logic, namely $S5$. It is used in spite of the fact that there are problems with these systems in specific situations. For example, axiom scheme (A5), $\neg K_i \phi \rightarrow K_i \neg K_i \phi$, does not hold for every person $i$ and statement $\phi$. Suppose someone does not know that the formalized version of L"ob’s theorems states that $\text{PA} \vdash \text{Prov}(\text{Prov}(\phi) \rightarrow \phi) \rightarrow \text{Prov}(\phi)$. As long as she does not hear or read anything about it, she will not know the existence of this theorem. Therefore, she does not know that she does not know what L"ob’s theorem states, so in this case (A5) does not hold. Similar to the fact that $S5$ is often accepted to represent knowledge, we accept, in spite of the mentioned problems, the interpretation of knowledge as provability.

5.3 Conclusion

The main question we want to answer in this thesis is the following. To what extent can provability logic be used to solve the knower paradox? A summary of the quality of the three systems which were discussed is presented in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>Skyrms</th>
<th>Anderson</th>
<th>Solovay</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Formal</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2. Philosophical</td>
<td>x</td>
<td>✓...</td>
<td>✓...</td>
</tr>
<tr>
<td>3. Scope</td>
<td>Not too narrow</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Not too broad</td>
<td>x</td>
<td>✓...</td>
</tr>
</tbody>
</table>

Table 2: Summary. The symbol ‘✓’ means ‘satisfied’, ‘x’ means ‘not satisfied’, and the addition of ‘...’ means ‘tentatively’.

We see that for Anderson’s solution and Solovay’s system, all of Haack’s requirements are at least provisionally satisfied. We added to Haack’s description of the requirement on the formal part of the solution that, besides a rejected premise or principle of inference, a rejected set of theorems could be indicated. We provisionally conclude that provability logic can be used to solve the knower paradox. It can turn out that it is not, if for both systems an example is found which proves that the systems are too broad as solutions to the knower paradox. In addition, the systems by Anderson and Solovay can appear to be failing to solve the paradox if some arguments are found that take down the argument of interpreting knowledge as provability and the argument of epistemic conservativity. This is the extent to which provability logic, as far as we considered interpretations of this, solves the knower paradox.
References


L. Åqvist. Deontic tense logic with historical necessity, frame constants, and a solution to the epistemic obligation paradox (the “knower”?). *Theoria*, 2014.


R. Dedekind. *Was sind und was sollen die Zahlen?* Friedr. Vieweg & Sohn, Braunschweig, 1888.


