Journal of Philosophical Logic (2023) 52:1101–1132
https://doi.org/10.1007/s10992-023-09699-3

Solutions to the Knower Paradox in the Light of Haack's Criteria

Mirjam de Vos[1] · **Rineke Verbrugge**[2] · **Barteld Kooi**[3]

Received: 23 December 2022 / Accepted: 2 January 2023 / Published online: 17 April 2023

Abstract

The knower paradox states that the statement 'We know that this statement is false' leads to inconsistency. This article presents a fresh look at this paradox and some well-known solutions from the literature. Paul Égré discusses three possible solutions that modal provability logic provides for the paradox by surveying and comparing three different provability interpretations of modality, originally described by Skyrms, Anderson, and Solovay. In this article, some background is explained to clarify Égré's solutions, all three of which hinge on intricacies of provability logic and its arithmetical interpretations. To check whether Égré's solutions are satisfactory, we use the criteria for solutions to paradoxes defined by Susan Haack and we propose some refinements of them. This article aims to describe to what extent the knower paradox can be solved using provability logic and to what extent the solutions proposed in the literature satisfy Haack's criteria. Finally, the article offers some reflections on the relation between knowledge, proof, and provability, as inspired by the knower paradox and its solutions.

Keywords Epistemic logic · Knower paradox · Provability logic · Haack's criteria

✉ Rineke Verbrugge
L.C.Verbrugge@rug.nl

[1] University of Groningen, Groningen, The Netherlands

[2] Department of Artificial Intelligence, Bernoulli Institute, Faculty of Science and Engineering, University of Groningen, Groningen, The Netherlands

[3] Department of Theoretical Philosophy, Faculty of Philosophy, University of Groningen, Groningen, The Netherlands

1 Introduction: The Knower Paradox

A paradox can be defined as "an apparently unacceptable conclusion derived by apparently acceptable reasoning from apparently acceptable premises" [37, p. 1]. This is the definition that we use throughout this article; for brevity's sake we will sometimes just state that a paradox is a certain "apparently unacceptable conclusion".

To set the stage, we first give an informal explanation of the knower paradox[1], after which we describe the formal version of the knower paradox as it was presented originally by Kaplan and Montague [21]. The knower paradox is based on the following statement:

$$\text{We know that statement } \mathrm{P} \text{ is false.} \tag{P}$$

Statement P is used to create the apparently unacceptable conclusion that 'P is true if and only if P is false', which is a paradox. We assume the principle of bivalence, which states that every statement is either true or false.

Suppose P is true. We assume that everything that is known is true[2]. Since statement P states that 'we know that statement P is false', it follows that statement P 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 P is true cannot be true. Because this is the case, we infer that statement P is false. Since we are the ones who *proved* that P is false, it follows that we *know* that P is false[3]. However 'we know that statement P 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, from this we inferred that it is false, which implies that it is true. This means that P is true if and only if P is false.

1.1 The Original Formalization of the Knower Paradox

For their 1960 formalization of the knower paradox, Kaplan and Montague [21] used *elementary syntax*, by which they understood "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" [21, Footnote 10, p. 89]. Note that by elementary syntax they meant both

[1] There are some paradoxes that go by names similar to the knower paradox with which the knower paradox should not be confused. One example is the knowledge or *knowability paradox* by Fitch [13]. This paradox of knowability is a logical result implying that, necessarily, if all truths are knowable in principle then all truths are in fact known.

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 the guard knows that the thing happens. Åqvist [2] 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" [24].

[2] It is a common assumption in epistemology that knowledge implies truth. As a reminder, Hintikka explains that it is "obvious that [this] condition has to be imposed on model sets" [19, p. 43]. The same principle is stated by Lenzen: "gewußt werden kann nur, was auch wahr ist" [22, p. 52]. Meyer an Van der Hoek introduce the axiom scheme $K_i\phi \rightarrow \phi$ as a property of knowledge [25, p. 23].

[3] By the principle that having a proof leads to knowledge, see for example [43].

a formal language and a proof system with axioms and derivation rules. Robinson's Arithmetic Q is a minimal formal system that has all elements Kaplan and Montague mention.

As a reminder, Robinson's arithmetic Q [36] is a formal theory extending first-order logic with identity. Its language $\mathcal{L}_A$ is built by induction from $0, S, +, \cdot, =$. The axioms of Q are the following.

$$\forall x(0 \neq Sx) \tag{1}$$
$$\forall x \forall y(Sx = Sy \rightarrow x = y) \tag{2}$$
$$\forall x(x \neq 0 \rightarrow \exists y(x = Sy)) \tag{3}$$
$$\forall x(x + 0 = x) \tag{4}$$
$$\forall x \forall y(x + Sy = S(x + y)) \tag{5}$$
$$\forall x(x \cdot 0 = 0) \tag{6}$$
$$\forall x \forall y(x \cdot Sy = (x \cdot y) + x) \tag{7}$$

A statement φ is a theorem of Q if it is (an instance of) an axiom or if it can be derived from the axioms in the sense that there exists "a sequence $\varphi_0, \ldots, \varphi_n$ of [formulae from $\mathcal{L}_A$] such that φ_n is φ and for each $i \leq n$, either φ_i is an axiom (...) or φ_i follows from some preceding members of the sequence using a rule of inference" [17, p. 7–8]. The available rules of inference are modus ponens and generalization [5, p. 19]. If statement φ is a theorem of Q, this is denoted by '$\mathrm{Q} \vdash \phi$'.

Kaplan and Montague used '$\varphi \vdash \psi$' to express that ψ is derivable from φ within the theory and '$\vdash \varphi$' means that φ is provable within this theory. In addition, they used names for expressions, where $\overline{\varphi}$ denotes the name of expression φ. These names can be defined via Gödel numbering [14]. Using this, it is possible to create self-referential arithmetical statements. The following two formulae are added to the elementary syntax:

$$K(\overline{\varphi}) \qquad A \text{ knows the expression } \varphi$$
$$I(\overline{\varphi}, \overline{\psi}) \qquad \varphi \vdash \psi$$

In modal multi-agent epistemic logic, $K_i\varphi$ is considered as a *sentential operator* K_i that can be applied to a sentence φ. A predicate $K(\overline{\varphi})$ with sentence name $\overline{\varphi}$ as argument is called a *metalinguistic predicate*. In both cases, the result of applying an operation to a sentence or applying a predicate to a term is a sentence. We consider the following statement: "A knows that the present statement is false". According to Kaplan and Montague [21, p. 87], we can regard some sentence D as expressing this statement, namely D satisfying

$$\vdash D \leftrightarrow K(\overline{\neg D}).$$

From this expression, some version of the knower paradox is derived, if the following three assumptions are made:

$$E1 := K(\overline{\neg D}) \rightarrow \neg D \tag{E1}$$
$$E2 := K(\overline{E_1}) \tag{E2}$$
$$E3 := [I(\overline{E_1}, \overline{\neg D}) \wedge K(\overline{E_1})] \rightarrow K(\overline{\neg D}) \tag{E3}$$

These premises are apparently acceptable. The assumption $E1$ says that if A knows the expression $\neg D$, then $\neg D$ is true. This corresponds to the idea that a falsehood cannot be known (see Footnote 2). Assumption $E2$ expresses that assumption $E1$ is known by A. It is a common assumption that A knows that what she knows is true, and $E2$ just expresses that this is the case for knowing $\neg D$. Finally, $E3$ expresses that if $\neg D$ is derivable from $E1$ and A knows $E1$, then A knows $\neg D$. This is an example of the epistemic closure principle: if $\vdash \varphi \rightarrow \psi$, then $\vdash K_i\varphi \rightarrow K_i\psi$. It is not an instance of axiom schema $(K_i\varphi \wedge K_i(\varphi \rightarrow \psi)) \rightarrow K_i\psi$, because $I(\overline{\varphi}, \overline{\psi})$ does not correspond to $K_i(\varphi \rightarrow \psi)$.

From these assumptions $E1$, $E2$ and $E3$, the knower paradox can be derived as the apparently unacceptable conclusion '$\vdash D \leftrightarrow \neg D$'. In the derivation of this, we use the notation of rules such as HS for Hypothetical Syllogism, MP for Modus Ponens and PC for Propositional Calculus. We denote '$Ei \vdash \varphi$' in proof line (j) if the definition of Ei is used to derive the statement in line (j) or a statement in one of the previous lines (1), (2), ..., $(j-1)$. In step (6) we use the following. If $\varphi \vdash \psi$, then $\vdash I((\overline{\varphi}, \overline{\psi}))$. By the diagonalization lemma [5, 14], it is shown that there exists a sentence D such that $D \leftrightarrow K(\overline{\neg D})$ is provable for D in the language L_A of Peano Arithmetic and $K(y)$ a formula of L_A in which no variable other than y is free. We derive the knower paradox as follows.

(1)	$\vdash D \leftrightarrow K(\overline{\neg D})$	by definition of D
(2)	$\vdash D \rightarrow K(\overline{\neg D})$	by (1), PC
(3)	$E1 \vdash K(\overline{\neg D}) \rightarrow \neg D$	by definition of $E1$
(4)	$E1 \vdash D \rightarrow \neg D$	by (2), (3), HS
(5)	$E1 \vdash \neg D$	by (4), PC
(6)	$E1 \vdash I(\overline{E_1}, \overline{\neg D})$	(5), by definition of I
(7)	$E1, E2 \vdash K(\overline{E1})$	by definition of $E2$
(8)	$E1, E2 \vdash I(\overline{E_1}, \overline{\neg D}) \wedge K(\overline{E_1})$	by (6), (7), PC
(9)	$E1, E2, E3 \vdash [I(\overline{E_1}, \overline{\neg D}) \wedge K(\overline{E_1})] \rightarrow K(\overline{\neg D})$	by definition of $E3$
(10)	$E1, E2, E3 \vdash K(\overline{\neg D})$	by (8), (9), MP
(11)	$E1, E2, E3 \vdash K(\overline{\neg D}) \rightarrow D$	by (1), PC
(12)	$E1, E2, E3 \vdash D$	by (10), (11), MP
(13)	$E1, E2, E3 \vdash \neg D \rightarrow D$	by (12), PC
(14)	$E1, E2, E3 \vdash D \leftrightarrow \neg D$	by (4), (13), PC

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 $E1$, $E2$, and $E3$ are made is inconsistent.

1.2 The Current Debate on the Knower Paradox

Even though the knower paradox has been introduced by Kaplan and Montague in 1960 and many solutions have been proposed, it is still the subject of heated debates, to which we now turn. There is only little consensus yet about how the knower paradox should be solved. The assumption that knowledge entails truth is accepted, of which $E1$ from Section 1.1 is an instance. There are ongoing debates about other parts of the paradox. Should the syntax be changed in such a way that statements that lead to paradoxes are eliminated? Should we accept the epistemic closure principle or not?

For example, Dean and Kurokawa [7, 8] write about a discussion between Cross [6] and Uzquiano [44]. The discussion is about the status of some assumption Cross uses in a version of the knower paradox which is slightly different from the original formulation. In the current article, we focus on two contributions to the debate about the knower paradox, both focusing on solutions that are based on provability logic and its variants, as well as various interpretations of these modal logics in formal systems of arithmetic.

The article that we discuss at length is Paul Égré's [11]. Égré argues that the knower paradox is solvable when modal provability logic is applied. He uses three different interpretations of provability logic to solve the paradox, namely interpretations by Skyrms [38], Anderson [1] and Solovay [42]. We also discuss Poggiolesi's [32]. Poggiolesi compares Anderson's and Solovay's solutions to the knower paradox and comments on Égré's solution, which she sees as an attempt to connect the first two. Our main contribution is an assessment of how the three interpretations by Skyrms, Anderson and Solovay fare in the light of Susan Haack's criteria for solutions to paradoxes [16], which include both technical and philosophical desiderata. In this way we hope to advance the debate regarding the knower paradox. In addition, we formulate an extension of Haack's criteria.

The rest of this article is structured as follows. In Section 2, we discuss Haack's criteria for solutions to paradoxes. We give a short reminder of provability logic and formal systems of arithmetic in Section 3, to set the stage for the discussion of the knower paradox. In Section 4, we explain the provability interpretations that Égré considers as solutions to the knower paradox. In addition, we discuss the quality of these solutions. We check to what extent they satisfy Haack's criteria and we evaluate whether some criticism on Égré's article by Poggiolesi is valid. In this way, we explain to what extent the knower paradox can be solved using provability logic.

2 Haack's Criteria for Solutions to Paradoxes

There is extensive literature on almost every paradox, yet often in this literature we find papers that lack any discussion on what actually constitutes a solution to a paradox. In her book *Philosophy of Logics* [16] Susan Haack offers general criteria for the solution of paradoxes. This is very worthwhile since it makes clear what is actually problematic when we are faced with a paradox and it provides a tool with which we can evaluate proposed solutions to paradoxes.

There are two different kinds of solutions to paradoxes. As a reminder, a statement is paradoxical if it is an apparently unacceptable conclusion derived by apparently acceptable reasoning from apparently acceptable premises [37]. A paradox is solved if:

1. we discard one of the axioms or rules of inference and accept the resulting theory in which the 'apparently unacceptable conclusion' cannot be derived;
2. in the new theory the conclusion can again be formulated but is not 'apparently unacceptable', which it was in the old system.

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' [33]. Many conclusions that are considered as paradoxical in other systems are not unacceptable in a dialetheic account. In this article, we focus on the first kind of solutions, in which the 'apparently unacceptable conclusion' cannot be derived.

Susan Haack describes three requirements on solutions to paradoxes. First, a solution should provide 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 that particular premise or principle of inference seems acceptable but is unacceptable. The third requirement is that a solution should not be too broad or too narrow.[4] We consider these requirements in more detail.

2.1 The Formal Part of a Solution (First Requirement)

According to Haack, 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)" [16, p. 138–139][5] Suppose we want to solve the liar paradox, then we need a consistent formal theory[6] Σ 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 article, 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. Note that this is only a minor adaptation of

[4]Recent applications of Haack's criteria to solutions of other semantic paradoxes can be found in [10, 20, 29].

[5]Note that for a paraconsistent logician the requirement should be about non-triviality rather than consistency. Switching to a paraconsistent view can solve certain paradoxes, but in the literature we are interested in consistency and non-triviality coincide, so we do not delve into this issue here.

[6]Note that a formal theory is not necessarily recursively axiomatized, for example Gupta's revision theory of truth [15] and Field's theory of truth and conditional [12].

Haack's ideas to the context of the knower paradox, that is fully in line with her general approach.

The system in which the paradox is formulated consists of a set of theorems, defined by premises and rules of inference. This system is defined in a certain language. By forming a new system, in which one of the premises or rules of inference from the old system is rejected or which is based on another language, we arrive at a new set of theorems. Except the 'apparently unacceptable conclusion', there might be other theorems that 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.

2.2 The Philosophical Part of a Solution (Second Requirement)

After stating some requirements to the formal solution to a paradox , Haack continues that a solution should "supply some explanation of *why* that premise or principle is, despite appearances, exceptionable (the *philosophical* solution)" [16, p. 139]. 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".

To continue the example above, 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. According to Haack, we need to find philosophical arguments for disallowing apparently acceptable principles of inference and premises in order to have a satisfactory solution.

2.3 The Scope of a Solution (Third Requirement)

A solution to a paradox is required to have the right scope, which means that it should be neither too broad nor too narrow. A solution is too broad if it is "so broad as to cripple reasoning we want to keep" [16, p. 139], and it is too narrow if it does not block all paradoxes that are closely related to the paradox under consideration. It is often somewhat vague which paradoxes are closely related to a given paradox. For example, if the solution solves a paradox of the form 'P if and only if $\neg P$', then should other paradoxes of this form be considered as closely related to it? Obviously not, but are the liar paradox and the knower paradox closely related because both involve self-reference? It may depend on the sort of solution that is proposed. If the solution revolves around an analysis of self-reference, one may consider them to be closely related. If the solution focusses on the concept of knowledge one may consider them to be unrelated.

Let us assume that in a certain context we are not bothered by this inherent vagueness and it is clear which group of paradoxes are to be solved. We can then explain the concept of scope 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. Note that S_2 may have a different language than S_1. 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 or its language would not even contain these statements. The solution is too broad if it rejects a statement from set A or if one of these statements is not translated into the language of the new system. The solution of a paradox is too narrow if one of the statements of set B is still in the language and still an acceptable conclusion. Note that a solution can be both too broad and too narrow.

To continue the running example, suppose that 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'.[7] 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.

In summary, 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 neither too broad nor too narrow. In Section 4, we evaluate three interpretations of provability logic as solutions to the knower paradox, using Haack's criteria as our yardstick.

3 Provability Logic and Formal Systems of Arithmetic

Before we consider some solutions to the knower paradox, we look at provability logic and formal systems of arithmetic. First we consider Peano arithmetic, after which we define a certain provability logic and its relation to arithmetic. In addition, we mention the diagonal lemma, which is used in the original formulation of the knower paradox.

3.1 Peano Arithmetic

Let us give a reminder of the most well-known extension of Robinson arithmetic. Peano arithmetic (PA) is named after Giuseppe Peano [31], who made a precise formulation of a set of axioms which had been proposed by Richard Dedekind [9]. To define PA, we need the following *Induction Schema*.

$$\{\varphi(0) \wedge \forall x(\varphi(x) \rightarrow \varphi(Sx))\} \rightarrow \forall x\varphi(x) \tag{8}$$

[7]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.

The axioms of PA are exactly all axioms of Q plus each instance of this induction schema. If statement φ is a theorem of PA, this is denoted by 'PA $\vdash \phi$'.

3.2 Provability Logic

The most widely used provability logic[8] is called **GL** and contains all axiom schemes from **K** and the extra scheme GL:

$$\text{All (instances of) propositional tautologies} \tag{A1}$$
$$\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi) \tag{A2}$$

$$\Box(\Box\varphi \to \varphi) \to \Box\varphi \tag{GL}$$

The rules of inference of **GL** are modus ponens and necessitation (if $\varphi \in \mathbf{GL}$, then $\Box\varphi \in \mathbf{GL}$). Note that $\Box\varphi \to \Box\Box\varphi \in \mathbf{GL}$ [45].

There are three conditions that a predicate $Prov(\overline{\varphi})$ should satisfy in order to be a *provability predicate* for arithmetical theory Σ.

$$\text{If } \Sigma \vdash \varphi \text{ then } \Sigma \vdash Prov(\overline{\varphi}) \tag{L1}$$
$$\Sigma \vdash Prov(\overline{\varphi \to \psi}) \to (Prov(\overline{\varphi}) \to Prov(\overline{\psi})) \tag{L2}$$
$$\Sigma \vdash Prov(\overline{\varphi}) \to Prov(\overline{Prov(\overline{\varphi})}) \tag{L3}$$

These conditions are called the Hilbert-Bernays-Löb derivability conditions[9] or just Löb's derivability conditions.[10] Löb proved that S, satisfying $\Sigma \vdash S \leftrightarrow Prov(\overline{S})$, is provable for $Prov(\overline{S})$ satisfying the derivability conditions [23]. This theorem can also be formulated as follows. If PA $\nvdash S$, then PA $\nvdash Prov(\overline{S}) \to S$. Gödel's second incompleteness theorem states that if PA $\nvdash \bot$, then PA $\nvdash \neg Prov(\overline{\bot})$ [14]. This can be proved from Löb's theorem.

3.3 The Relation between Provability Logic and Peano Arithmetic

Note that the derivability conditions for PA correspond to the principles of **GL**. To make this more precise, we now describe the important relation between formal arithmetic PA and provability logic **GL**, using the definition of a realization. A *realization* is a function that assigns to each propositional atom of modal logic a sentence of the language of arithmetic. The inductive definition of the realization $*$ is given by the following clauses.

$$\begin{aligned} \bot^* &= \bot \\ (\varphi \to \psi)^* &= (\varphi^* \to \psi^*) \\ (\Box\varphi)^* &= Prov(\overline{\varphi^*}) \end{aligned}$$

[8] Instead of **GL**, this logic is sometimes called **KW**, **KW**, **K4W**, **PrL** or **L**.
[9] See [5, p. 16], [39, p. 223].
[10] See [41, p. 118], [18, 45].

Other logical connectives like $(\varphi \wedge \psi)$ can be defined by $\rightarrow$ and $\bot$, so $*$ also respects these. This definition of realization $*$ is used in the defintion of arithmetical soundness and completeness. In 1976, Robert Solovay [42] proved that **GL** is arithmetically complete with respect to PA. The arithmetical soundness of **GL** was already clear. So **GL** is arithmetically complete ('if') and arithmetically sound ('only if') w.r.t. Peano arithmetic, which means that

$$\mathbf{GL} \vdash \varphi \text{ if and only if PA } \vdash \varphi^* \text{ for all realizations } * .$$

So **GL** "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" [45].

We now consider the diagonal lemma[11], which makes it possible to introduce certain self-referential sentences. It proves for example that statement D, which is used to provide the original knower paradox in Section 1.1, can indeed be defined. Statement D satisfies $\Sigma \vdash D \leftrightarrow K(\overline{\neg D})$. The diagonal lemma is stated as follows.

Theorem 1 (Diagonal Lemma, [5, p. 54][12]) *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 $\vdash S \leftrightarrow P(\overline{S})$.*

A clear sketch of the proof can be found in a supplement of an article by Raatikainen [35].

4 Solutions to the Knower Paradox in the Light of Provability Logic

We try to solve the knower paradox using provability logic. In Section 3.2, we discussed some theorems by Gödel and Löb, which play an important role in this logic. As Visser 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" [46, p. 793]. 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 that solve the knower paradox according to Égré [11]. In addition, we discuss whether these solutions are satisfactory by discussing some articles that commented on them and by applying Haack's requirements, described in Section 2, to the solutions.

4.1 Different Treatments of Modalities as Used in Solutions to the Knower Paradox

Before we consider the solutions that Égré describes, we define four kinds of treatments of modalities, namely sentential treatments on the one hand, and

[11] Smith calls it 'Diagonalization Lemma' and explains that it deserves the status of being a theorem rather than being a lemma [39, p. 173].

[12] We replace Boolos' '$\ulcorner S \urcorner$' by '$\overline{S}$'.

metalinguistic, syntactical and arithmetical treatments on the other. Like in the first section of this article, a *sentential* operator applies to sentences, but a *metalinguistic* predicate applies to names of sentences. If some metalinguistic predicate is self-referential, such as a predicate to which the diagonal lemma applies, then we call it *syntactical*. Finally, an *arithmetical* predicate is a specific kind of 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 these four different kinds of operators are shown in Fig. 1.

Important in Égré's article is that a syntactical treatment, defined by Montague [27] and Cross [6] 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 with self-reference*, as explained by Égré [11, p. 34]. In addition, Égré shows that there exist both a consistent non-metalinguistic treatment with self-reference and a consistent metalinguistic treatment of modalities which is not self-referential.

As a first interpretation of provability logic used to solve the knower paradox, Égré mentions a theory by Skyrms [38] as a consistent *metalinguistic treatment* of modalities which does not contain self-referential statements like knower sentences D (satisfying $D \leftrightarrow K(\overline{\neg D})$). In contrast, Égré describes two examples of *self-referential* systems. In a nutshell, Anderson's system [1] weakens the axiom scheme $K(\overline{K(\overline{\varphi}) \rightarrow \varphi})$. Solovay's system [42] weakens the necessitation rule of inference to prevent that scheme $K(\overline{K(\overline{\varphi}) \rightarrow \varphi})$ results from applying necessitation to the axiom scheme $K(\overline{\varphi}) \rightarrow \varphi$.

We discuss the systems by Skyrms, Anderson and Solovay in Sections 4.2, 4.3 and 4.4 respectively. Although only Anderson published his system with the goal to contribute to the discussion about the knower paradox, Égré explains that all three of the theories provide solutions to the knower paradox. We consider the quality of these solutions in the light of Haack's criteria.

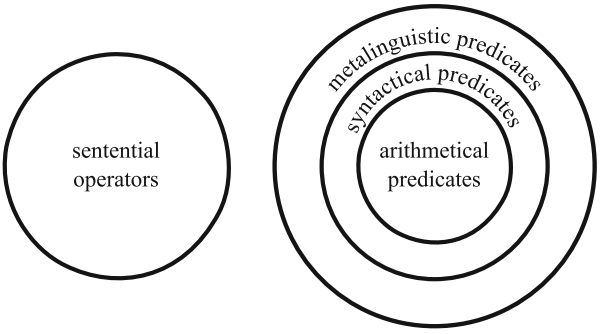

Fig. 1 The relation between the different kinds of operators as used in treatments of modalities

4.2 Skyrms's Interpretation of Provability Logic

We consider Skyrms's interpretation of provability logic [38] and discuss why Égré [11] states that this is a solution to the knower paradox. Skyrms himself does not mention the knower paradox in his article.

By the derivation of the original knower paradox in Section 1.1, we know that arithmetical treatments of modalities can lead to inconsistencies. Égré explains that Skyrms shows that there does exist a consistent form of metalinguistic treatment of modalities. Suppose $\Box\varphi$ is metalinguistically interpreted as 'φ is provable'. Skyrms defines modal language L_M as follows, where L_0 is a finitary language containing the language of the propositional calculus:

$$\begin{aligned} L_M: \quad & \text{A language containing } L_0, \\ & \text{closed under Boolean operators,} \\ & \text{for which } \varphi \in L_M \text{ implies } \Box\varphi \in L_M \end{aligned}$$

The counterpart of L_M is based on the same language L_0 and is defined by induction:

$$\begin{aligned} L_0: \quad & \text{A finitary language containing propositional calculus.} \\ L_{n+1}: \quad & \text{The smallest extension of } L_n \text{ such that if } \varphi \in L_n, \text{ then } Prov(\text{`}\varphi\text{'}) \in L_{n+1}, \\ & \text{closed under Boolean operators.} \\ L_\omega = \quad & \bigcup_{n\in\omega} L_n \end{aligned}$$

The predicate $Prov(\text{`}\varphi\text{'})$ expresses that $\varphi$ 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'. Égré only considers the provability interpretation and writes $*Q(S)$ as $Prov(\text{`}\varphi\text{'})$. Skyrms explains that "[t]he expression consisting of a sentence prefixed by 'Q' is to be thought of as a name for that sentence" [38, p. 369–370]. So 'φ' in $Prov(\text{`}\varphi\text{'})$ does not express the numeral corresponding to the name of φ, but expresses just the name of φ. This means that the treatment of modalities in Skyrms's system is metalinguistic. It is not syntactical, since it does not contain self-referential statements in which the predicate $Prov$ occurs.

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

$$\begin{aligned} t(\varphi) &= \varphi && \text{for all } \varphi \in L_0 \\ t(\Box\varphi) &= Prov(\text{`}t(\varphi)\text{'}) && \text{for all } \varphi \in L_M \\ & t \text{ distributes over the truth-functional connectives} \end{aligned}$$

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

Why is Skyrms's system, consisting of an hierarchy of languages and a translation from L_M to L_ω, a solution to the knower paradox? It is a consistent theory of language L_ω, which does not contain sentences of the form $D \leftrightarrow Prove(\text{`}\neg D\text{'})$. To show this, Égré states the following consistency result.

Theorem 2 (Consistency of Skyrms's System [11, 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 $\varphi \in T_n$, then $Prov(\ulcorner\varphi\urcorner) \in T_{n+1}$*
2. *If $\varphi, \psi \in L_n$, then $Prov(\ulcorner\varphi \to \psi\urcorner) \to (Prov(\ulcorner\varphi\urcorner) \to Prov(\ulcorner\psi\urcorner)) \in T_{n+1}$*
3. *If $\varphi \in L_n$, then $Prov(\ulcorner\varphi\urcorner) \to \varphi \in T_{n+1}$*
4. *If $Prov(\ulcorner\varphi\urcorner) \in L_n$, then $Prov(\ulcorner\varphi\urcorner) \to Prov(\ulcorner Prov(\ulcorner\varphi\urcorner)\urcorner) \in T_{n+1}$*

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

According to Égré, who gives a short proof of this theorem, "[t]his consistency result shows that the theory T_ω, although it is an extension of Robinson [a]rithmetic, can satisfy all the metalinguistic translations of the modal schemata involved in (...) the weak system T-Nec used to present the [knower paradox]" [11, p. 37]. In contrast to T-Nec, Skyrms's system treats the predicate $Prov$ in a way that does not contain self-referential sentences like $D \leftrightarrow Prov(\ulcorner\neg D\urcorner)$. This leads to the difference in consistency of the systems.

Why is $D \leftrightarrow K(\overline{\neg D})$ not contained in Skyrms's system, meaning that there is no D of the form $D \leftrightarrow Prov(\ulcorner\neg D\urcorner) \in L_\omega$, satisfying $D \leftrightarrow Prov(\ulcorner\neg D\urcorner) \in T_\omega$? There is some form of self-refence in Skyrms's system, because it extends weak arithmetic, but it does not interfere with the predicate $Prov$. Égré states that "[t]he core of Skyrms's approach is indeed to sever the self-referential apparatus of arithmetic from the metalinguistic system used to handle the predicate $Prov$" [11, p. 37–38]. We will show that there is no D of the form $D \leftrightarrow Prov(\ulcorner\neg D\urcorner)$ that is in L_ω and such that $D \leftrightarrow Prov(\ulcorner\neg D\urcorner) \in T_\omega$ (assuming Robinson arithmetic Q is consistent). We prove this by contradiction.

Suppose there is some D of the form $D \leftrightarrow Prov(\ulcorner\neg D\urcorner)$ in L_ω, satisfying $D \leftrightarrow Prov(\ulcorner\neg D\urcorner) \in T_\omega$. Then there exists some $n \in \omega$ such that $D \leftrightarrow Prov(\ulcorner\neg D\urcorner) \in T_n \subset T_\omega$. This means that $D \to Prov(\ulcorner\neg D\urcorner) \in T_n$. Since $D \in L_\omega$, there exists an m for which $D \in L_m$. Then $\neg D \in L_m$, since L_m is closed under Boolean operators. Therefore, by requirement 3 of Theorem 2, $Prov(\ulcorner\neg D\urcorner) \to \neg D \in T_m \subset T_\omega$ follows. By deductive closure from $D \to Prov(\ulcorner\neg D\urcorner) \in T_\omega$ and $Prov(\ulcorner\neg D\urcorner) \to \neg D \in T_\omega$, we have $D \to \neg D \in T_\omega$. This implies $\neg D \in T_\omega$ by propositional logic. By requirement 1 of Theorem 2, $Prov(\ulcorner\neg D\urcorner) \in T_\omega$ follows. From the original assumption, we derive $Prov(\ulcorner\neg D\urcorner) \to D \in T_\omega$. Then $D \in T_\omega$ by deductive closure. Since both $\neg D \in T_\omega$ and $D \in T_\omega$, T_ω is inconsistent. This contradicts with Theorem 2, stating that T_ω is consistent if Q is consistent. We conclude that there is no D of the form $D \leftrightarrow Prov(\ulcorner\neg D\urcorner)$ in L_ω such that $D \leftrightarrow Prov(\ulcorner\neg D\urcorner) \in T_\omega$.

The first step of the original derivation of the paradox consisted of '$\vdash D \leftrightarrow K(\overline{\neg D})$' (see Section 1.1, Page 3). Although the derivation by Kaplan and Montague resembles the one described above in a certain way, there is a crusial difference. In the original derivation, the existence of such a sentence D followed from the diagonalization lemma. We do not have this in Skyrms's system, since $\vdash D \leftrightarrow K(\overline{\neg D})$ only holds for T_{n+1} if $D \in T_n$ and not $\vdash D \leftrightarrow K(\overline{\neg D}) \in T_n$.

Since $D \leftrightarrow Prov(\text{`}\neg D\text{'}) \notin T_\omega$, the knower paradox cannot be derived in Skyrms's system T_ω in the same way as we did in Section 1.1. Therefore, accepting T_ω solves the knower paradox. We discuss the extent to which this solution satisfies the requirements by Haack [16] in Sections 4.2.1, 4.2.2 and 4.2.3.

4.2.1 The Formal Part of Skyrms's Theory as a Solution

As we discussed above, Skyrms [38] proposes to treat modalities in a metalinguistic way without self-referential statements of a certain form. No knower sentence is contained in Skyrms's system T_ω, so the knower paradox cannot be derived in the original way, described in Section 1.1.

Is Skyrms's 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 above, T_ω is consistent if Q is consistent.[13] Besides, theorems like $D \leftrightarrow Prov(\text{`}\neg D\text{'})$ are not in this new theory which describes knowledge. This means that Skyrms's system satisfies Haack's first criterion as a solution to the knower paradox.

4.2.2 The Philosophical Part of Skyrms's Theory as a Solution

Does Skyrms's 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 $D \leftrightarrow K(\overline{\neg D})$ in the theory that describes knowledge. The article by Skryms [38] is about modalities in general, but not specifically about knowledge. It starts with a reference to Quine [34], who takes the view that the most natural construal of modalities is as predicates applying to names of sentences, so as metalinguistic predicates . This is an argument for treating modalities metalinguistically, but not for disallowing $D \leftrightarrow K(\overline{\neg D})$. We can at least appreciate that Skyrms's motivation is independent of the paradox.

In addition, Skyrms [38, 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". Skyrms's provability interpretation leads to an interpretation of **S4**. This means that the modal principles which hold for language L_0, defined as finitary language containing propositional calculus, are exactly the principles of **S4**. So arguments for accepting **S4** as a system to describe knowledge are also arguments for accepting Skyrms's system, but this does not directly indicate why we should disallow self-referential statements like $D \leftrightarrow K(\overline{\neg D})$. So we do not see arguments for disallowing the rejected set of theorems, which means that Haack's second requirement is provisionally not satisfied.

[13] It is generally assumed by mathematicians that Robinson arithmetic Q, being a sub-theory of PA, is indeed consistent.

4.2.3 The Scope of Skyrms's Theory as a Solution

Haack's third requirement states that a solution to a paradox should not be too broad or too narrow. A solution which is consistent satisfies the requirement that it should not be too narrow, because consistency implies that we have not ended up with a different paradox, such as the liar paradox. As we have seen in Section 4.2.1, Skyrms's system is consistent. So this system is not too narrow.

However, Skyrms's system T_ω is too broad as a solution to the knower paradox, because it does not contain some non-paradoxical statement such as fixed-point statements. Thus, it throws the baby out with the bathwater, as we will proceed to show. A Gödel equivalence G for T_ω, with $G \leftrightarrow \neg Prov(\text{`}G\text{'})$, is such a statement. This sentence is relevant for a solution to the knower paradox, because it is a self-referential sentence about provability.

We show by contradiction that $G \leftrightarrow \neg Prov(\text{`}G\text{'})$ is not contained in $T_\omega$ for all $G \in L_\omega$ satisfying $G \leftrightarrow \neg Prov(\text{`}G\text{'})$. Suppose that $G \leftrightarrow \neg Prov(\text{`}G\text{'}) \in T_\omega = \cup_{n\in\omega} T_n$, then there exists some $n \in \omega$ such that $G \leftrightarrow \neg Prov(\text{`}G\text{'}) \in T_n$. This means that $G \rightarrow \neg Prov(\text{`}G\text{'}) \in T_n$. Since $G \in L_\omega$, there exists an $m$ for which $G \in L_m$. By requirement 3 of Theorem 2, $Prov(\text{`}G\text{'}) \rightarrow G \in T_m \subset T_\omega$ follows. By deductive closure from $Prov(\text{`}G\text{'}) \rightarrow G \in T_\omega$ and $G \rightarrow \neg Prov(\text{`}G\text{'}) \in T_\omega$, we have $Prov(\text{`}G\text{'}) \rightarrow \neg Prov(\text{`}G\text{'}) \in T_\omega$. This implies $\neg Prov(\text{`}G\text{'}) \in T_\omega$ by propositional logic. From the original assumption, we derive $\neg Prov(\text{`}G\text{'}) \rightarrow G \in T_\omega$. Since $\neg Prov(\text{`}G\text{'}) \in T_\omega$, it follows that $G \in T_\omega$. By requirement 1 of Theorem 2, $Prov(\text{`}G\text{'}) \in T_\omega$ follows. Since both $\neg Prov(\text{`}G\text{'}) \in T_\omega$ and $Prov(\text{`}G\text{'}) \in T_\omega$, T_ω is inconsistent. This contradicts with Theorem 2, stating that T_ω is consistent if Q is consistent. We conclude that there is no $G \in L_\omega$ satisfying $G \leftrightarrow \neg Prov(\text{`}G\text{'})$ such that $G \leftrightarrow \neg Prov(\text{`}G\text{'}) \in T_\omega$, so Skyrms's system T_ω misses all fixed-point sentences with respect to Skyrms's provability predicate. This means that Skyrms's system as a solution to the knower paradox is too broad.[14]

We conclude that Skyrms's 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's system is not too narrow, but it is too broad.

4.3 Anderson's Interpretation of Provability Logic

Let us now consider Anderson's provability interpretation of epistemic logic and discuss why both Anderson and Égré state that this provides a solution to the knower paradox.

Skyrms's system as a solution to the knower paradox abandoned a certain form of self-reference in his theory T_ω. Anderson [1] argues that we should not abandon self-reference, but modify the incompatible axiom schemes that lead to the

[14]Note that the Gödel sentence with arithmetical provability predicate Pr_Q for Q itself is contained in Q and thus in T_ω.

paradox. Anderson considers the following three generalizations of the axioms $E1$, $E2$, and $E3$ from the original knower paradox by Kaplan and Montague [21] (see Section 1.1):

$$K(\overline{\varphi}) \rightarrow \varphi \tag{T}$$

$$K(\overline{K(\overline{\varphi}) \rightarrow \varphi}) \tag{U}$$

$$[I(\overline{\varphi}, \overline{\psi}) \wedge K(\overline{\varphi})] \rightarrow K(\overline{\psi}) \tag{I}$$

As we will see, Anderson constructs a hierarchy with 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[15], where L_A is the language of Robinson and Peano arithmetic.

$$\begin{aligned}
L_0 \;:&\; \text{the smallest extension of } L_A \text{ such that} \\
&\text{if } \varphi,\ \psi \in L_A, \text{ then } K_0(\overline{\varphi}),\ I_0(\overline{\varphi}, \overline{\psi}) \in L_0, \\
&\text{closed under Boolean operators.} \\
L_{i+1} \;:&\; \text{the smallest extension of } L_i \text{ such that} \\
&\text{if } \varphi,\ \psi \in L_i, \text{ then } K_{i+1}(\overline{\varphi}),\ I_{i+1}(\overline{\varphi}, \overline{\psi}) \in L_{i+1}, \\
&\text{closed under Boolean operators.} \\
L_\omega \;=&\; \bigcup_{i \in \omega} L_i
\end{aligned}$$

Notice that this K_i does not mean 'agent i knows', but indicates a certain *level* of knowledge. Anderson gives an "intuitive motivation", inspired by John Myhill [28], for accepting more than one knowledge predicate [1, p. 348–349]. The idea is as follows. Some sentence that cannot be in a set of statements known at level i can still be provable. By understanding the proof of such a statement, one knows this sentence at level $i + 1$.

It is assumed that there is a given Gödel numbering for L_ω, 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:

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

[15] Égré [11, 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 [1, p. 351–352] himself states that "language L_i is obtained from L_ω 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}.

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:

$$\begin{aligned} T_0 &= Q \cup \{K_0(\overline{\varphi}) \to \varphi \mid \varphi \in L_\omega\} \\ T_{i+1} &= T_i \cup \{K_{i+1}(\overline{\varphi}) \to \varphi \mid \varphi \in L_\omega\} \\ V_0(K_0(\overline{\varphi})) &= 1 \text{ if and only if } Q \vdash \varphi \\ V_{i+1}(K_{i+1}(\overline{\varphi})) &= 1 \text{ if and only if } T_i \vdash \varphi \\ V_0(I_0(\overline{\varphi}, \overline{\psi})) &= 1 \text{ if and only if } Q \vdash \varphi \to \psi \\ V_{i+1}(I_{i+1}(\overline{\varphi}, \overline{\psi})) &= 1 \text{ if and only if } T_i \vdash \varphi \to \psi \end{aligned}$$

In this article, we consider axiom set $T_\omega = \cup_{i\in\omega} T_i$ 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 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 $n = gn(\varphi) \in V_i(K_i)$, then $\exists j \geq i$ such that $V_j(\varphi) = 1$.
If $n = gn(\varphi)$, $m = gn(\psi)$, $(n, m) \in V_i(I_i)$, then $\exists j \geq i$ such that $V_j(\varphi \to \psi) = 1$.
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:

$$\begin{aligned} V(K_i(\overline{\varphi}) \to \varphi) &= 1 \\ V([I_i(\overline{\varphi}, \overline{\psi}) \wedge K_i(\overline{\varphi})] \to K_i(\overline{\psi})) &= 1 \\ V(K_{i+1}(\overline{K_i(\overline{\varphi}) \to \varphi})) &= 1 \end{aligned}$$

By the first two of these statements, we still have T and I in Anderson's system. There are two different forms of U, namely $K_{i+1}(\overline{K_i(\overline{\varphi}) \to \varphi})$ and $K_i(\overline{K_i(\overline{\varphi}) \to \varphi})$. The first one is valid, but if we use this one in the derivation of the knower paradox as described in Section 1.1 on Page 3, then we will not arrive at an inconsistency. This is the case, because we get $K_{i+1}(\overline{\neg D})$ in Step (10) of the derivation and $K_i(\overline{\neg D}) \to D$ in Step (11), which does not give us D. Therefore, we cannot conclude the inconsistency of D with $\neg D$. Applying the other form, $K_i(\overline{K_i(\overline{\varphi}) \to \varphi})$, would lead to the inconsistency in the same way as described in Section 1.1 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, $T_j \vdash K_i(\overline{\varphi}) \to \varphi$ holds only for $j \geq i$. This means that $T_{i-1} \vdash K_i(\overline{\varphi}) \to \varphi$ does not hold, so by definition of interpretation V_i, $V_i(K_i(\overline{K_i(\overline{\varphi}) \to \varphi})) \neq 1$. 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. Let us consider the extent to which Anderson's solution satisfies Haack's requirements.

4.3.1 The Formal Part of Anderson's Solution

In Anderson's hierarchy of languages, the formula $K_i(\overline{K_i(\overline{\varphi}) \rightarrow \varphi})$ is rejected, which implies that no formula representing the knower paradox can be derived in the way that was shown in Section 1.1. Does this mean that the first requirement from Haack [16] 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 $K(\overline{K(\overline{\varphi}) \rightarrow \varphi})$. Is Anderson's theory also consistent? Dean and Kurokawa [8, p. 221] write about the consistency proof that Anderson sketches for his theory. The statement "$V_0(K_0(\overline{\varphi})) = 1$ if and only if $Q \vdash \varphi$" holds (step (i)). The rest of the proof sketch is then formulated as "$V_{i+1}(K_{i+1}(\overline{\varphi})) = 1$ if and only if $T_i \vdash \varphi$" for $T_0 = Q \cup \{K_0(\overline{\varphi}) \rightarrow \varphi \mid \varphi \in L_\omega\}$ and $T_{i+1} = T_i \cup \{K_{i+1}(\overline{\varphi}) \rightarrow \varphi \mid \varphi \in L_\omega\}$. 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 $\neg\psi \in \{K_0(\overline{\varphi}) \rightarrow \varphi \mid \varphi \in L_\omega\}$, or $\psi \in \{K_0(\overline{\varphi}) \rightarrow \varphi \mid \varphi \in L_\omega\}$ 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.[16] So Anderson's solution meets Haack's first requirement.

4.3.2 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 $K(\overline{K(\overline{\varphi}) \rightarrow \varphi})$? Using articles by Anderson [1] and Poggiolesi [32], we arrive at an argument for rejecting U.

The following argument to disallow axiom scheme U is given both by Anderson [1, p. 350] and Poggiolesi [32, p. 152]. The axiom sceme U 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 $K(\overline{\varphi}) \rightarrow \varphi$, $K(\overline{K(\overline{\varphi}) \rightarrow \varphi})$, and $[I(\overline{\varphi}, \overline{\psi}) \wedge K(\overline{\varphi})] \rightarrow K(\overline{\psi})$ (T, U, and I respectively). Presuming a provability interpretation, the schemes $Prov(\overline{\varphi}) \rightarrow \varphi$ and $[I(\overline{\varphi}, \overline{\psi}) \wedge Prov(\overline{\varphi})] \rightarrow Prov(\overline{\psi})$ are valid, while U, interpreted as $Prov(\overline{Prov(\overline{\varphi}) \rightarrow \varphi})$, is not. So interpreting knowledge as provability implies that U should be disallowed. The connection between knowledge and provability is further discussed in Section 5.2.

Poggiolesi [32] argues that Anderson's intuitive argument for introducing different knowledge levels fails, because it uses two different notions of proof. She claims that "there is no reason for changing the notion of proof on which (...) knowledge is based" [32, p. 157]. We do not agree with Poggiolesi here that the use of different notions of proof would be 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 (which are not formalizable in the system K_0),

[16] We assume that Robinson arithmetic is consistent.

it is plausible to define at least two different kinds, or levels, of knowledge. As an analogy, consider the sequence of mathematical theories:

1. PA;
2. $PA + Con(PA)$.
3. $PA + Con(PA + Con(PA))$, etc.

Based on Gödel's second completeness theorem, it is immediately clear that these theories differ from one another. For example, if PA is consistent, then the second level proves $Con(PA)$, which is true but not provable at the first level, and so on. To us, provability in these theories does intuitively correspond to increasing levels of knowledge, thereby saving the philosophical part of Anderson's solution.

The argument that axiom scheme U is not valid in a system where provability is considered as knowledge forms the philosophical part of Anderson's solution to the knower paradox. This philosophical part indicates objections to the rejected principle U. The first reason to reject U is because it is not valid if we interpret knowledge as provability. We think this argument forms 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.

4.3.3 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. Anderson's solution to the knower paradox is consistent, so just as Skyrms's solution, Anderson's solution is not too narrow.

A statement that is potentially able to show that a solution to a paradox is too broad is Gödel sentence G, which satisfies $G \leftrightarrow \neg K_i(\overline{G})$ for some $i \in \omega$. Suppose that $T_{i=j}$ is the first theory of Anderson's hierarchy in which G occurs. Then $T_{j-1} \nvdash G$, so $V_j(K_j(\overline{G})) = 0$. Therefore, we have $V_j(G) = 1$ and $V_j(\neg K_j(\overline{G})) = 1$, so $V_j(G \leftrightarrow \neg K_j(\overline{G})) = 1$. This means that there is indeed some $i \in \omega$, namely $i = j$, for which $G \leftrightarrow \neg K_i(\overline{G}) \in T_i$. 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.

4.3.4 Reflecting on Haack's Criteria

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 the argument by Poggiolesi [32] that we explained in Section 4.3.2, satisfies all of Haack's requirements on solutions to paradoxes at least provisionally. Still, the idea of more than one level of knowledge has not yet been motivated independently of the paradox. Does this mean that Haack's criteria are not sufficient for

assessing the quality of a solution to a paradox? It seems that a philosophical solution that only explains why a premise or principle is to be disallowed, does not by itself provide a good story for accepting a different premise or principle that is to replace a problematic one. When we are faced with multiple solutions to a paradox that all reject the same premise, surely a satisfactory solution would also have to provide an argument why the new premise is better than its alternatives. Indeed the implicit premise that Anderson rejects ('there is exactly one knowledge predicate') is also rejected by Dean and Kurokawa, yet their solutions differ significantly in the number of knowledge predicates that replace the single knowledge predicate in the system of Kaplan and Montague. So maybe a general requirement to the solution of paradoxes should be added to Haack's list, namely the requirement that philosophical motivation should be provided for a new premise or principle that replaces a rejected premise or principle.

Égré [11, p. 40] states that "[t]he strength of [Anderson's] solution, as compared to [Skyrms's 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 [42], as solution to the knower paradox, which according to Égré [11, p. 38] has a "significant connection" with Anderson's system.

4.4 Solovay's Interpretation of Provability Logic

We discussed Skyrms's consistent system, in which there is one provability predicate and self-referential sentences cannot be proved in T_ω. We also considered Anderson's hierarchy of languages, in which infinitely many provability predicates occur but self-referential sentences can be valid. According to Égré [11, 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 $\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$ and $\Box(\Box\varphi \to \varphi) \to \Box\varphi$, and the inference rules modus ponens and necessitation (see Section 3.2). The system **GLS**, defined by Solovay [42, Section 5.1][17], contains all theorems of **GL** as axioms as well as all instances of the reflection principle $\Box\varphi \to \varphi$, 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 [42].

Why is the knower paradox prevented in **GLS**? Remember that $K(\overline{E1})$, where $E1$ was defined as $K(\overline{\neg D}) \to \neg D$, was needed in the derivation of the knower paradox by Kaplan and Montague [21] (see Section 1.1, Page 3, Step (7)). In **GLS**, we have $\Box\neg D \to \neg D$ as an instance of the reflection principle. Because necessitation is not an inference rule of **GLS**, $\Box(\Box\neg D \to \neg D)$ cannot be derived from the reflection principle here, therefore, Kaplan and Montague's derivation cannot be repeated in **GLS**.

[17] We follow current conventions as in e.g. [5, 45] in that Solovay's G is our **GL** and his G' is our **GLS**.

Égré states more about Solovay's system, in particular about its connection to the one by Anderson. We cite[18] 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]. [11, p. 43]

By instances of the reflection principle, Égré means all instances of $Prov(\overline{\varphi}) \rightarrow \varphi$, where $Prov$ means provability in PA. Note that PA^+ is not closed under necessitation. Égré makes the following four claims. The first is that **GLS** corresponds to PA^+. The second claim is that PA^+ can prove the consistency of PA. As the last two claims, Égré states that PA^+ is the counterpart of T_0 in Anderson's system (described in Section 4.3) and that T_0 contains an implicit hierarchy. We discuss these four claims consecutively.

(1) Why does GLS Correspond to PA^+? We think Égré means that **GLS** corresponds to PA^+ in the same way as **GL** corresponds to PA. The system **GL** is arithmetically sound and arithmetically complete with respect to PA, which means that $\mathbf{GL} \vdash \varphi$ if and only if $\mathrm{PA} \vdash \varphi^*$ for all realizations $*$. Is it the case that $\mathbf{GLS} \vdash \varphi$ if and only if $\mathrm{PA}^+ \vdash \varphi^*$ for all realizations $*$, so we can say that **GLS** corresponds to PA^+?[19]

Solovay [42, Section 5.1] proves that **GLS** is arithmetically sound and arithmetically complete with respect to the standard model $\langle \omega; +, \cdot \rangle$. In addition, PA^+ is sound with respect to this standard model, so $\mathrm{PA}^+ \vdash \varphi$ implies $\omega \models \varphi$. So if $\mathbf{GLS} \nvdash \varphi$, then by the completeness part of Solovay's theorem some realization $*$ exists such that $\omega \not\models \varphi^*$. By soundness of PA^+ with respect to ω, it follows that $\mathrm{PA}^+ \nvdash \varphi^*$. So assuming $\mathbf{GLS} \nvdash \varphi$, it follows that $\mathrm{PA}^+ \nvdash \varphi^*$ for some realization $*$. This means that **GLS** is arithmetically complete with respect to PA^+.

To prove the arithmetical soundness of **GLS** with respect to PA^+, the arithmetical soundness of **GL** with respect to PA can be used. We also use a theorem from Boolos [5, p. 131], according to which $\mathbf{GLS} \vdash \varphi$ implies that there exist $\psi_1, \ldots \psi_n$ such that $\mathbf{GL} \vdash \bigwedge\{\Box\psi_i \rightarrow \psi_i \mid i = 1, \ldots, n\} \rightarrow \varphi$. Since **GL** is sound with respect to PA, it follows that $\mathrm{PA} \vdash \bigwedge\{Prov(\overline{\psi_i^*}) \rightarrow \psi_i^* \mid i = 1, \ldots, n\} \rightarrow \varphi^*$ for all realizations $*$. Because PA^+ contains all instances of $Prov(\overline{\psi_i^*}) \rightarrow \psi_i^*$, we conclude that $\mathrm{PA}^+ \vdash \varphi^*$ for all realizations $*$. This means that **GLS** is arithmetically sound with respect to PA^+.

[18] We use our own notation of **GLS**, PA, T, etc.

[19] For the answer to this question, personal communication with Paul Égré himself was used, for which we are very grateful.

We conclude that **GLS** is arithmetically complete and arithmetically sound with respect to PA^+. Therefore, **GLS** corresponds to PA^+.

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

(3) Why is PA^+ the Counterpart of T_0 in Anderson's System? First we need to know what it means that PA^+ is the counterpart of T_0. We consider an article by Poggiolesi [32], who explains that there are two ways to interpret the correspondence of **GLS** with PA^+. She argues that both interpretations are incorrect because they imply $\mathrm{PA}^+ \neq T_0$. We don't think that $\mathrm{PA}^+ = T_0$ is meant by stating that "PA^+ is the counterpart of T_0". Égré [11, p. 26] also talks about T', U' and I' as counterparts of T, U, and I (described in Section 4.3), where in T', U' and I', K is replaced by K' as the knowledge-plus predicate[20] defined by Cross [6]. Thus T is the axiom scheme $K(\overline{\varphi}) \rightarrow \varphi$ and T' is $K'(\overline{\varphi}) \rightarrow \varphi$. It is not the case that T$=T'$, so we think Égré also does not mean to say that $\mathrm{PA}^+ = T_0$. In addition, Égré [11, 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 equivalent to this other scheme. We think PA^+ being the counterpart of T_0 means that PA^+ and T_0 contain only axioms which are one another's counterparts, like the axiom schemes $\Box\varphi \rightarrow \varphi$ from PA^+ and $K_0(\overline{\varphi}) \rightarrow \varphi$ 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 $[K(\overline{\varphi}) \wedge I(\overline{\varphi}, \overline{\psi})] \rightarrow K(\overline{\psi})$ while PA^+ does not. PA^+ does contain $[K(\overline{\varphi}) \wedge K(\overline{\varphi \rightarrow \psi})] \rightarrow K(\overline{\psi})$, but these two schemes are "only equivalent (...) in the presence of the translation of the rule of necessitation, that is not, as we already said, a rule of PA^+" [32, p. 161]. We agree with Poggiolesi that the schemes are not equivalent, but that does not mean that they cannot be counterpart of each other. In particular, the epistemic closure principle is stronger than the scheme in PA^+.

We think that PA^+ is the counterpart of T_0, because PA^+ extends PA in an *analogous* way to how T_0 extends Q. To arrive at T_0 from Q, all instances of $K_0(\overline{\varphi}) \rightarrow \varphi$ are added for $\varphi \in L_\omega$. To arrive at PA^+ from PA, all instances of $Prov(\overline{\varphi}) \rightarrow \varphi$ are added for φ in the language of PA^+, and modus ponens is applied. Like PA, Q is closed under modus ponens. Since only instances of $K_0(\overline{\varphi}) \rightarrow \varphi$ 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 PA^+ is an extension of PA in the same way as T_0 is an extension of Q, we conclude that PA^+ can be seen as counterpart of T_0.

[20] Cross defined $K'(\overline{x})$ as $\exists y(K(\overline{y}) \wedge I(\overline{y}, \overline{x}))$.

(4) What is the Implicit Hierarchy that is present within T_0 of Anderson's System? Anderson defined $T_0 = Q \cup \{K_0(\overline{\varphi}) \rightarrow \varphi \mid \varphi \in L_\omega\}$ and $T_{i+1} = T_i \cup \{K_{i+1}(\overline{\varphi}) \rightarrow \varphi \mid \varphi \in L_\omega\}$ (for $i \in \omega$, $i \neq 0$). Alternatively, he could have defined $T_0 = Q$ and $T_{i+1} = T_i \cup \{K_i(\overline{\varphi}) \rightarrow \varphi \mid \varphi \in L_\omega\}$. 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 Égré, and these point to a similarity between Anderson's T_0 and Solovay's **GLS**. In Solovay's system, we do not have $\Box(\Box\varphi \rightarrow \varphi)$, because the necessitation rule is not applied to instances of the reflection principle $\Box\varphi \rightarrow \varphi$. Do we have something like this for Anderson's system? T_0 contains all instances of $K_0(\overline{\varphi}) \rightarrow \varphi$, just like **GLS** contains all instances of $\Box\varphi \rightarrow \varphi$ and PA^+ contains all instances of $Prov(\overline{\varphi}) \rightarrow \varphi$. 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 $K_0(\overline{K_0(\overline{\varphi}) \rightarrow \varphi})$. The only kind of necessitation that can be applied in Anderson's system to arrive at something like U, is a rule which concludes $K_{n+1}(\overline{K_n(\overline{\varphi}) \rightarrow \varphi}) \in T_{n+1}$ from $K_n(\overline{\varphi}) \rightarrow \varphi \in T_n$. This one does not result in an instance of U that can be used to derive the knower paradox.

Égré [11, 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 [21]. 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.

Égré [11, p. 43] adds two last sentences before his concluding remarks. "In **GLS**, the [necessitation rule] allows to iterate schemata **K** and **4** arbitrarily many times. But the reflection principle [T] cannot be iterated systematically, thereby preventing the appearance [of] the [k]nower paradox." Here, **K** is the axiom scheme $K_n(\varphi \rightarrow \psi) \rightarrow (K_n\varphi \rightarrow K_n\psi)$ and **4** is the scheme $K_n\varphi \rightarrow K_nK_n\varphi$. If we consider **GLS** as a set of theorems for which only the inference rule modus ponens holds, this seems incorrect. However, if Égré considers **GLS** as a system containing the axioms of **GL** together with $\Box\varphi \rightarrow \varphi$ for which the necessitation rule only applies to the axioms of **GL** and modus ponens to all axioms, then the contents of the quotation is correct. In **GL**, necessitation can be applied to **K** and **4**, but in **GLS**, there is no necessitation rule available that can be applied to reflection principle T. Let us now asses to which extent Solovay's theories satisfy Haack's criteria for solutions to paradoxes.

4.4.1 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 $K(\overline{K(\overline{\varphi}) \rightarrow \varphi})$, which is achieved by disallowing the necessitation rule to apply to the reflection principle $K(\overline{\varphi}) \rightarrow \varphi$. Is **GLS** consistent? Solovay [42] proved that **GLS** is arithmetically sound with respect to the standard

model. Since truth in a model implies consistency, **GLS** is consistent. So Haack's first requirement on solutions to paradoxes is satisfied.

4.4.2 The Philosophical Part of Solovay's Theory as a Solution

To satisfy Haack's second requirement, there needs to be an argument for rejecting $K(\overline{K(\overline{\varphi}) \rightarrow \varphi})$ or for disallowing the necessitation rule to apply to the reflection principle $K(\overline{\varphi}) \rightarrow \varphi$. This argumentation should be independent of the existence of the knower paradox. Solovay [42] did not consider **GLS** within the context of the knower paradox. His article is about provability and not about knowledge, so we do not find arguments for rejecting $K(\overline{K(\overline{\varphi}) \rightarrow \varphi})$ there. Considering provability, there are reasons to reject $Prov(\overline{Prov(\overline{\varphi}) \rightarrow \varphi})$. Löb's theorem states that PA $\vdash Prov(\overline{Prov(\overline{\varphi}) \rightarrow \varphi}) \rightarrow Prov(\overline{\varphi})$. This implies that if $Prov(\overline{Prov(\overline{\varphi}) \rightarrow \varphi})$ is accepted as an axiom scheme, then $Prov(\overline{\varphi})$ holds for every statement φ, 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.

Égré [11, 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, $K(\overline{\varphi}) \rightarrow \varphi$, 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 [11] 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 provisionally 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.

4.4.3 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's and Anderson's system (see Sections 4.2 and 4.3), we conclude that a system is not too narrow if it is consistent. Solovay's system is consistent, so it is not too narrow.

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 consider the same example as in Sections 4.2.3 and 4.3.3. Gödel sentence G in PA satisfies PA $\vdash G \leftrightarrow \neg Prov(\overline{G})$. 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 it is an instantiation of the reflection principle. The formula $\neg\Box\bot \leftrightarrow \neg\Box\neg\Box\bot$ is in **GL** (as an instance of De Jongh and Sambin's fixed-point theorem for provability logic; for a proof, see [45, Section 2.2]), and thus in **GLS**. Since $\Box(\Box\bot \rightarrow \bot) \rightarrow \Box\bot$ is an axiom of **GL**, it follows that **GL** $\vdash \Box(\neg\Box\bot) \rightarrow \Box\bot$. So there is some G, namely

Table 1 Summary of Section 4. The symbol '✓' means 'satisfied', 'x' means 'not satisfied', and the addition of '. . . ' means 'provisionally'

		Skyrms	Anderson	Solovay
1. Formal		✓	✓	✓
2. Philosophical		x. . .	✓. . .	x. . .
3. Scope	Not too narrow	✓	✓	✓
	Not too broad	x	✓. . .	✓. . .

$\neg\Box\bot$, which satisfies $\mathbf{GLS} \vdash G \leftrightarrow \neg\Box G$, which means that a Gödel sentence is a theorem of Solovay's system.[21] 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.

4.5 Summary

In this section, we explained the three different solutions to the knower paradox described by Égré [11]. The different solutions reject different parts of the derivation of the knower paradox by Kaplan and Montague [21] (see Section 1.1, Page 3). Skyrms abandons the validity of the statement $D \leftrightarrow K(\overline{\neg D})$ and thereby rejects the first step of the derivation. Anderson's solution 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 article is to explain to what extent the knower paradox can be solved using provability logic. We discussed the quality of the theories of Skyrms [38], Anderson [1] and Solovay [42]. Consider Table 1 for a summary of this discussion.

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 an argument by Poggiolesi [32], but provisionally not by Skyrms's 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 argument.

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

[21] This can alternatively be shown by the interesting fact that **GL** can be alternatively axiomatized without the modalized Löb axiom GL but as "diagonalization logic". This is the modal logic **S4** plus the new rule: From $((p \leftrightarrow A(p)) \wedge \Box(p \leftrightarrow A(p))) \rightarrow B$, derive B, where all occurrences of p occur under the scope of $\Box$ in $A(p)$ and p does not occur in B [40, Theorem 2.5]. We would like to thank one of the anonymous reviewers for pointing out this alternative proof.

in Skyrms's system, but it is in Anderson's and Solovay's systems. So Skyrms's 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 Closing Remarks

We want to answer the following question. To what extent can provability logic be used to solve the knower paradox? In this final section we consider an improvement of one of the solutions discussed in Section 4 and we comment on the idea of interpreting knowledge as provability in general.

5.1 Trying to Improve Égré's Solutions

In Section 4, three systems that represent provability were used by Égré to interpret knowledge. We discussed to what extent these solutions satisfied the requirements by Haack, described in Section 2. The solution by Anderson [1] satisfies all these requirements at least provisionally, while the solutions by Skyrms [38] and Solovay [42] do not satisfy the requirement on the philosophical part of the solution. In this section, we discuss an improvement of the solution that uses Solovay's system and compare this to Anderson's solution.

Improving the Philosophical Part of Solovay's System as a Solution In Section 4.4, we noted that Solovay's solution did not satisfy Haack's second requirement, that required arguments for disallowing the rejected premise, principle of inference, or set of theorems.

In Section 4.3.2, we described a reason by which Anderson's system satisfies Haack's second requirement. This argument to disallow axiom scheme U can also be used to complete Égré's idea to use Solovay's system as a solution to the knower paradox.

Just like for Anderson's solution, accepting an interpretation of provability as knowledge is a good reason to accept Solovay's system as a solution to the knower paradox. Solovay's **GLS** is a system about provability which is arithmetically complete and arithmetically sound with respect to the standard model ω. This indicates that **GLS** describes mathematical knowledge, namely facts about provability in Peano arithmetic which are known by mathematicians.

We add a second argument to disallow axiom scheme U in **GLS**. Solovay's system, **GLS**, is epistemically conservative over PA, meaning that **GLS** will not prove any 'new' formulas of the form 'It is known that φ', i.e. $\Box\varphi$, for which Peano Arithmetic does not prove φ^* yet (cf. [8]). We can see this by the following argument. Since **GLS** is arithmetically sound with respect to the standard model $\langle\omega; +, \cdot\rangle$, $\mathbf{GLS} \vdash \Box\varphi$ implies $\omega \models Prov_{\mathrm{PA}}(\overline{\varphi^*})$ for all realizations $*$. This means that there exists a proof of φ^* in PA for all realizations $*$, so $\mathrm{PA} \vdash \varphi^*$ holds for every realization $*$. So for $\Box$ interpreted as knowledge, **GLS** is 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 Égré's idea to use Solovay's system **GLS** as a formal solution to the knower paradox. Therefore, we now conclude that this 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 Section 4 was that the interpretation of Anderson [1] is the best of these three, because it best meets the requirements on solutions to paradoxes by Haack [16]. We have found arguments which satisfy the philosophical part of Solovay's system as a 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. In Section 4.3, we mentioned Anderson's intuitive motivation for accepting more than one knowledge predicate. Anderson's use of different kinds of proofs could be an argument for the different knowledge levels. However, only two kinds of proofs are used in Anderson's reasoning, while an infinite number of knowledge predicates occurs in his system. So we do 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 do 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 bi-modal logics are discussed for example by Beklemishev [3] and Smoryński [40, Chapter 4].

Dean and Kurokawa [8] consider the search for even more provability predicates, which represent provability in many different axiomatic systems like Q, $I\Delta_0+EXP$, and extensions of PA. Each different provability predicate could be used as an interpretation of 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 $Prov_{\mathrm{Q}}$, for example by the definition of Hájek and Pudlák, knowledge of statements in $I\Delta_0 + EXP$ as $Prov_{I\Delta_0+EXP}$, for example by the definition of Hájek and Pudlák, 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 did add some arguments to Solovay's system 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 Égré [11] to interpret knowledge, applying a certain definition of provability.[22] Each of the three solutions contains provability in a theory which extends Robinson arithmetic. In Skyrms's system, $Prov(\ulcorner\varphi\urcorner)$ means 'φ is provable in T_ω'. In Anderson's system, $K_i(\overline{\varphi})$ means 'φ is known at level i', which is the case for $i = 0$ if φ is provable in Q′. In Solovay's system, $\Box\varphi$ means 'φ 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 intuitively plausible to interpret knowledge as provability. If a mathematician has a proof of some statement, then this person *knows* the proved statement. Thus, provability seems to imply knowledge. One could argue that the converse also holds. A statement can only be mathematical knowledge if it is also provable. If some statement about (Peano) arithmetic is not provable, then there is no proof of it, so no mathematician can know the statement.

However, there are also arguments against interpreting knowledge as provability. 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 that Platonists can use against interpreting knowledge as provability is that knowledge seems to be dependent on mathematicians and on time, while provability does not.

We stated that the existence of a proof implies that there is a person who came up with it. According to this non-Platonistic view, 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, at least according to Platonism.

[22] Only the second system, by Anderson [1], was used to interpret knowledge in the article in which it was published.

Technically, there are statements that are known but not provable in PA. There are also statements that 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 $\mathrm{PA} \vdash G \leftrightarrow \neg Prov(\overline{G})$. 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[23], whose truth can be shown in second-order arithmetic, but of which the Paris-Harrington theorem states that it is not provable in PA [30].

An example of a theorem which was provable in PA but not known, can be found by considering a theorem which had been a conjecture for some time and finally has been proved in PA: Catalan's conjecture. This conjecture states that the unique solution[24] in the natural numbers to $x^m - y^n = 1$ is $x = 3$, $y = 2$, $m = 2$, $n = 3$. While the conjecture was stated in 1844, a full proof was first given by Mihãilescu in 2002 [26]. This proof is partly based on logarithmic forms and electronic computations, but in 2005, Bilu [4] 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. This means that we have an example of something that 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 content was mathematical knowledge.

Another example of a theorem is Löb's theorem. The formalized version of this theorem, $\mathrm{PA} \vdash Prov(\overline{Prov(\overline{\varphi}) \rightarrow \varphi}) \rightarrow Prov(\overline{\varphi})$, 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 [5, 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 such as 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.

5.3 Conclusion

The main question we set out to answer in this article is: 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.

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

[23] 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.

[24] Assuming m, n are integers greater than 1 and x, y are both unequal to 0.

Table 2 Summary. The symbol '✓' means 'satisfied', 'x' means 'not satisfied', and the addition of '...' means 'provisionally'

		Skyrms	Anderson	Solovay
1. Formal		✓	✓	✓
2. Philosophical		x...	✓...	✓...
3. Scope	Not too narrow	✓	✓	✓
	Not too broad	x	✓...	✓...

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 also suggested that a requirement should be added which requires philosophical reasons to accept premises or theorems that replace rejected premises or theorems.

We provisionally conclude that provability logic can be used to solve the knower paradox. It can turn out that it is not[25], 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 fail 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 interpretations of provability logic solve the knower paradox.

Acknowledgements We would like to thank Paul Égré for illuminating e-mail discussions about his paper on the knower paradox. We are also grateful to Marc Pauly for proofreading an earlier version of this article. This work was partially supported by the Netherlands Organisation for Scientific Research (NWO) Vici grant NWO 277-80-001, awarded to Rineke Verbrugge for the project 'Cognitive systems in interaction: Logical and computational models of higher-order social cognition'. Finally, we would like to thank the anonymous reviewers and the editor Frank Veltman for their helpful suggestions which led to many important improvements, as well as for their patience.

Author Contributions BK, RV and MdV devised the research question and the research strategy. MdV did most of the initial research and made all pictures and tables. BK, RV and MdV all wrote parts of the original submitted manuscript, its two revisions, and the replies to reviewers.

Funding This work was partially supported by the Netherlands Organisation for Scientific Research (NWO) Vici grant NWO 277-80-001, awarded to Rineke Verbrugge for the project 'Cognitive systems in interaction: Logical and computational models of higher-order social cognition'.

[25] At least not for the three provability interpretations we discussed.

References

1. Anderson, C. A. (1983). The paradox of the knower. *The Journal of Philosophy*, *80*(6), 338–355.
2. Åqvist, L. (2014). Deontic tense logic with historical necessity, frame constants, and a solution to the epistemic obligation paradox (the "knower"). *Theoria*.
3. Beklemishev, L. D. (1994). On bimodal logics of provability. *Annals of Pure and Applied Logic*, *68*(2), 115–159.
4. Bilu, Y. F. (2005). Catalan without logarithmic forms (after Bugeaud, Hanrot and Mihăilescu. *Journal de Théorie des Nombres de Bordeau*, *17*(1), 69–85.
5. Boolos, G. (1995). *The logic of provability*. Cambridge: Cambridge University Press.
6. Cross, C. B. (2001). The paradox of the knower without epistemic closure. *Mind*, *110*(438), 319–333.
7. Dean, W. (2014). Montague's paradox, informal provability, and explicit modal logic. *Notre Dame Journal of Formal Logic*, *55*(2), 157–196.
8. Dean, W., & Kurokawa, H. (2014). The paradox of the Knower revisited. *Annals of Pure and Applied Logic*, *165*(1), 199–224.
9. Dedekind, R. (1888). *Was sind und was sollen die Zahlen?*. Braunschweig: Friedr Vieweg & Sohn.
10. Douven, I. (2005). A principled solution to Fitch's paradox. *Erkenntnis*, *62*(1), 47–69.
11. Égré, P. (2005). The knower paradox in the light of provability interpretations of modal logic. *Journal of Logic, Language and Information*, *14*(1), 13–48.
12. Field, H. (2008). Saving truth from paradox. Oxford University Press.
13. Fitch, F. B. (1963). A logical analysis of some value concepts. *Journal of Symbolic Logic*, *28*, 135–142.
14. Gödel, K. (1931). On formally undecidable propositions of Principia Mathematica and related systems I. In K. Gödel (Ed.) *Collected Works: Publications 1929–1936* (pp. 144–195). New York: Oxford University Press. Original publication: 1931; Collected Works: 1986.
15. Gupta, A., & Belnap, N. (1993). The revision theory of truth. MIT Press.
16. Haack, S. (1978). *Philosophy of logics*. Cambridge: Cambridge University Press.
17. Hájek, P., & Pudlák, P. (1993). *Metamathematics of First-Order Arithmetic*. Perspectives in Mathematical Logic. Springer Milan, 1993 Second Printing 1998 of the First Edition.
18. Halbach, V., & Visser, A. (2014). The Henkin sentence. In *The Life and Work of Leon Henkin* (pp. 249–263). Springer.
19. Hintikka, J. (1962). *Knowledge and belief: An Introduction to the Logic of the Two Notions, volume 181 of Contemporary Philosophy*. Ithaca and London: Cornell University Press.
20. Johnston, C. (2014). Conflicting rules and paradox. *Philosophy and Phenomenological Research*, *88*(2), 410–433.
21. Kaplan, D., & Montague, R. (1960). A paradox regained. *Notre Dame Journal of Formal Logic*, *1*(3), 79–90.
22. Lenzen, W. (1980). *Glauben, Wissen und Wahrscheinlichkeit: Systeme der Epistemischen Logik* Vol. 12. Wien - New York: Springer-Verlag. Library of Exact Philosophy.
23. Löb, M. H. (1955). Solution of a problem of Leon Henkin. *Journal of Symbolic Logic*, *20*(2), 115–118.
24. McNamara, P. (2014). Deontic logic. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy Spring 2014 edition*.
25. Meyer, J.-J. C. H., & van der Hoek, W. (2004). *Epistemic Logic for AI and Computer Science, volume 41 of Cambridge Tracts in Theoretical Computer Science*. Cambridge: Cambridge University Press.
26. Mihăilescu, P. (2004). Primary cyclotomic units and a proof of Catalan's conjecture. *Journal für die Reine und Angewandte Mathematik (Crelles Journal)*, *572*, 167–195.
27. Montague, R. (1963). Syntactical treatments of modality, with corollaries on reflexion principles and finite axiomatizability. *Acta Philosophica Fennica*, *16*, 153–167.
28. Myhill, J. (1960). Some remarks on the notion of proof. *The Journal of Philosophy*, *57*(14), 461–471.
29. Papazian, M. (2012). Chrysippus confronts the liar: the case for stoic cassationism. *History and Philosophy of Logic*, *33*(3), 197–214.
30. Paris, J., & Harrington, L. (1977). A mathematical incompleteness in Peano arithmetic. *Studies in Logic and the Foundations of Mathematics*, *90*, 1133–1142.
31. Peano, G. (1889). *Arithmetices Principia, Nova Methodo Exposita*. Turin: Fratres Bocca.
32. Poggiolesi, F. (2007). Three different solutions to the knower paradox. *Annali del Dipartimento di Filosofia*, *13*(1), 147–163.

33. Priest, G., & Berto, F. (2013). Dialetheism. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy, Summer 2013 edition*.
34. Quine, W. V. O. (1953). Three grades of modal involvement. In *The Ways of Paradox and Other Essays* (pp. 156–174). Cambridge: Harvard University Press.
35. Raatikainen, P. (2014). Gödel's incompleteness theorems. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy Winter 2014 edition*.
36. Robinson, R. M. (1950). An essentially undecidable axiom system. In *Proceedings of the International Congress of Mathematicians*, (Vol. 1 pp. 729–730). Cambridge.
37. Sainsbury, R. M. (2009). *Paradoxes*, 3rd. New York 1987: Cambridge University Press.
38. Skyrms, B. (1978). An immaculate conception of modality or how to confuse use and mention. *The Journal of Philosophy*, *75*(7), 368–387.
39. Smith, P. (2007). *An Introduction to Gödel's Theorems*. New York: Cambridge University Press.
40. Smoryński, C. (1985). *Self-Reference and Modal Logic*. New-York: Springer-Verlag.
41. Smoryński, C. (1991). The development of self-reference: Löb's theorem. In T. Drucker (Ed.) *Perspectives on the History of Mathematical Logic,* (pp. 110–133). *Springer*.
42. Solovay, R. M. (1976). Provability interpretations of modal logic. *Israel Journal of Mathematics*, *25*(3-4), 287–304.
43. Sorensen, R. (2014). Epistemic paradoxes. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy Spring 2014 edition*.
44. Uzquiano, G. (2004). The paradox of the knower without epistemic closure?. *Mind*, *113*(449), 95–107.
45. Verbrugge, R. (2017). Provability logic. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy Fall 2017 edition*.
46. Visser, A. (1998). Provability logic. In E. Craig (Ed.) *Routledge Encyclopedia of Philosophy*, (Vol. 7 pp. 793–797). Taylor & Francis.