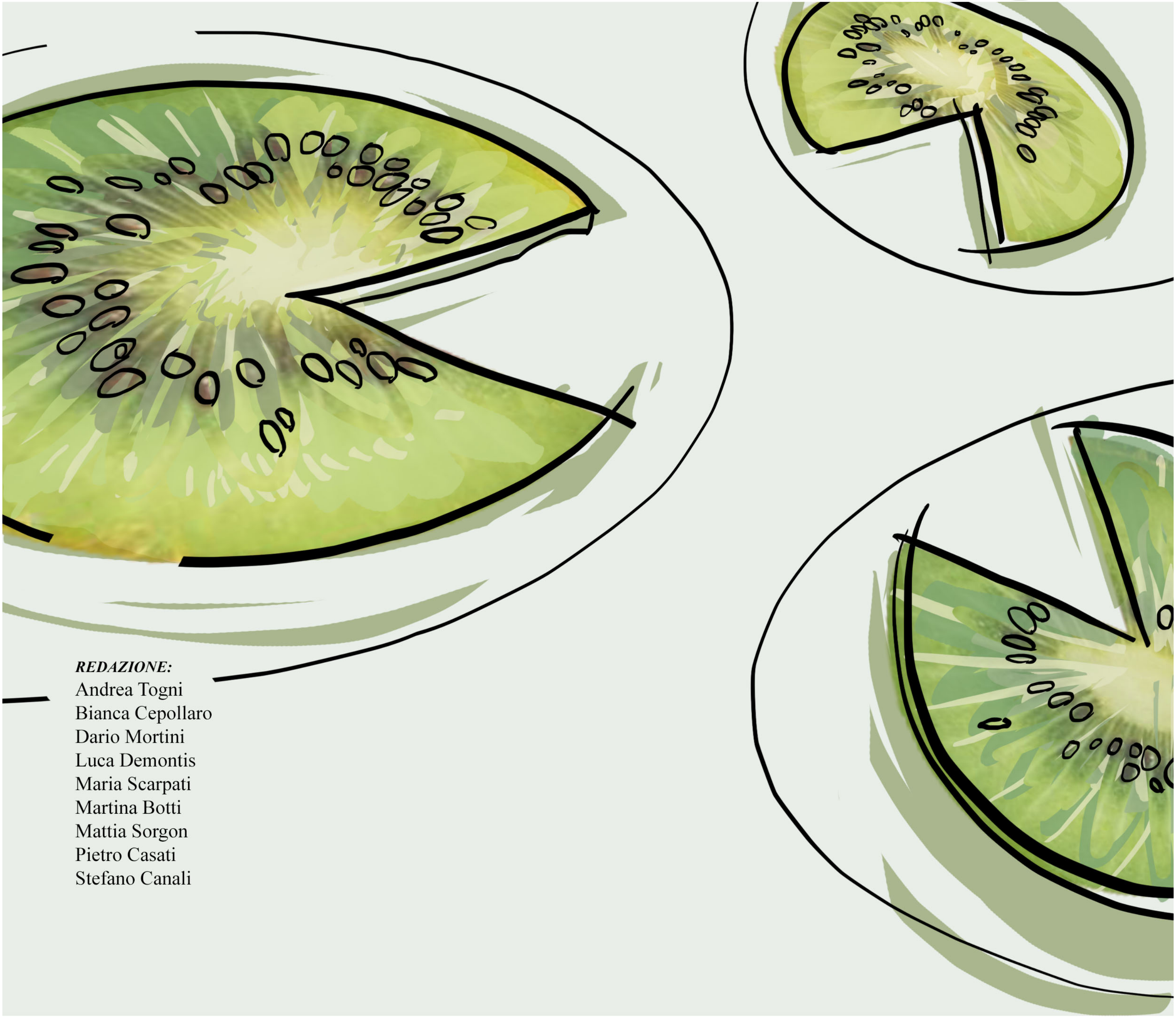
www.rifanalitica.it

Peer Review Journal

Rivista Italiana di Filosofia Analitica junior

VOLUME **7**
NUMERO **2**
2016



Patrocinata dalla
Società Italiana di Filosofia
Analitica dal 2011

Peer Review Journal
Rivista Italiana di Filosofia Analitica junior

ISSN: 2037 - 4445

Patrocinata dal 2011 dalla
Società Italiana di Filosofia Analitica

ISSN 2037-4445 ©
http://www.rifanalitica.it

Rivista Italiana di Filosofia Analitica Junior
Vol. 7, n. 2

Indice

Editoriale

Articoli

Patrocinata dal 2011 dalla
Società Italiana di Filosofia Analitica

ISSN 2037-4445 ⓒ
http://www.rifanalitica.it

Lavori in corso

Stefano Canali, Pietro Angelo Casati

Cari lettori, è con piacere che apriamo la pubblicazione di RIFAJ Vol.7 N.2. Si tratta di un'uscita un po' particolare, legata ad alcune novità riguardo al futuro del progetto RIFAJ.

In sei anni di attività la Rivista ha mantenuto una presenza costante, pubblicando puntualmente due numeri all'anno, ampliando il raggio delle iniziative correlate e coinvolgendo un maggior numero di collaboratori sia nella Redazione sia nel Comitato Scientifico. Ha dato a tanti la possibilità di mettersi in gioco in prima persona e per la prima volta. Molti di noi sono passati da laurea triennale a magistrale e dottorato, spostandosi nel frattempo in giro per (quasi) tutto il mondo.

Data la stabilità raggiunta, pensiamo sia giunto il momento per dedicarci allo sviluppo di alcune potenzialità inespresse, così da portare il progetto ad uno stadio successivo. Nei prossimi mesi RIFAJ andrà incontro ad una profonda rivisitazione, volta ad una crescita in termini di qualità scientifica, maturità e attività correlate. In quest'ottica abbiamo deciso di pubblicare il presente e il prossimo fascicolo con un numero ridotto di contributi, così da investire tempo ed energie nella ristrutturazione del progetto. Sarà perciò assente il consueto contorno di recensioni, interviste e report, nonché *Firma d'Autore* ed *Ex-Cathedra*, solitamente presenti nei nostri numeri tematici. In particolare, il volume corrente presenta due articoli, accomunati dal fatto di proporre revisioni di prospettive inaugurate da Saul Kripke.

In "Epistemic Logic and the Problem of Epistemic Closure", Davide Emilio Quadrellaro presenta una logica della conoscenza alternativa alle logiche proposizionali modali di tipo kripkeano. L'intento dell'autore è di servirsi dei "mondi impossibili" per evitare il compromesso con il controverso principio di chiusura epistemica.

In "A New Model for the Liar", Luca Castaldo presenta un nuovo modello per il linguaggio dell'aritmetica di Peano con l'aggiunta del predicato unario di verità. Estendendo il punto fisso minimo di Kripke e impiegando una particolare logica a quattro valori, l'autore intende sopperire ad un'inadeguatezza del modello

Autori. Stefano Canali, stefanocanali@me.com. Pietro Angelo Casati, pietroangelo.casati@gmail.com.

kripkeano, che non consente di distinguere tra i cosiddetti mentitori e quelli che potremmo chiamare assertori.

Entrambi gli articoli vertono dunque sulla Logica, in linea con quello che era stato presentato come uno Special Issue di Logica. Tuttavia, dato l'alto numero di potenziali contributi, la ricchezza del dibatto e la quantità e qualità di eventi e pubblicazioni sul tema, crediamo sia necessario costruire un numero tematico più corposo. Ci fa quindi piacere anticipare fin da subito che stiamo lavorando alla costruzione di un nuovo Special Issue di Logica, con uscita prevista a novembre 2017, per cui a breve diffonderemo un Call for Papers and Reviews. In attesa delle grandi novità, invitiamo ad inviare contributi, sottoporre nuove proposte e leggere i prossimi numeri.

Cogliamo infine l'occasione per ringraziare tutti coloro che ci hanno scritto, letto e supportato in questi sei anni.

Restate sintonizzati[1].

[1] A proposito di "sintonizzazione", vi invitiamo a seguirci sulla pagina Facebook di RIFAJ (facebook.com/RifanaliticaJun), che grazie all'amministrazione di Dario Mortini è decisamente più attiva che in precedenza e diventa sempre più ricca di segnalazioni, link notevoli, interviste variegate, spunti interessanti, nonché dilettevoli meme.

Sponsored since 2011 by the
Italian Society for Analytic Philosophy

ISSN 2037-4445 ⓒ
http://www.rifanalitica.it

Epistemic Logic and the Problem of Epistemic Closure

Davide Emilio Quadrellaro[*]

Abstract. This paper argues that propositional modal logics based on Kripke-structures cannot be accepted by epistemologists as a minimal framework to describe propositional knowledge. In fact, many authors have raised doubts over the validity of the so-called *principle of epistemic closure*, which is always valid in normal modal logics. This paper examines how this principle might be criticized and discusses one possible way to obtain a modal logic where it does not hold, namely through the introduction of impossible worlds..

Keywords. Epistemology, Epistemic Logic, Epistemic Closure, Rantala Semantics, Logic of Knowledge, Impossible Worlds.

[*] I wish to thank Greg Sax, Gianmarco Defendi and three anonymous referees for comments and suggestions.

Author. Davide Emilio Quadrellaro, davide.quadrellaro@gmail.com.
Received. 18 August 2016. **Accepted**. 19 October 2016.

1 Introduction

The purpose of this article is to describe a minimal logic of knowledge which can be used by epistemologists with different philosophical orientations. A first way to proceed is describing a modal logic based on a Kripke-semantics, specifying how the accessibility relation should be restricted in order to represent knowledge. However, it is not difficult to prove that this standard formal epistemological analysis implies the validity of the principle of epistemic closure, namely of the fact that, if one both knows that p and that if p then q, then he/she also knows that q. This principle, however, has been object of criticism and objections by some epistemologists. Therefore, if we are looking for a general modal logical framework that can be used by philosophers with different orientations, we have to construct a formal system where the closure principle does not hold. An interesting way to proceed is working with the semantics which has been developed by logicians to account for the paradox of the logical omniscience. In fact, if we introduce the "impossible worlds" and we construct a Rantala-semantics based on them, we obtain a weaker logic where the closure principle does not hold.

In the first part of this article I present the modal logic **T**, which is generally considered the minimal formal system for the logic of knowledge. Firstly I introduce the syntax and the semantics of modal logic, secondly I characterize how the accessibility relation R_a has to be restricted in order to obtain the logic **T**. In the second part I prove that the principle of epistemic closure follows from **T** and I try to underline some critical aspects of it. In the third part I introduce an alternative logic for knowledge where the closure principle does not hold, namely a modal logic with impossible worlds and a Rantala-semantics. Finally, in the fourth part, I evaluate this proposal, trying to underline both upsides and downsides of it.

2 The standard logic of knowledge

A first way to give a formal account to epistemological concepts such as belief and knowledge is to adopt the language of modal logic. Even if the modal operators $\Diamond$ and $\Box$ are usually read as possibility and necessity, we can also adopt an epistemic interpretation of them. On this alternative reading we will translate a logical formula like $\Box p$ not as "it is necessary that p" but rather as "it is known that p", "it is believed that p" or "it is certain that p". Following each of these interpretations we can formulate a different modal logic, in order to formalize the specific features of the considered epistemic operator. In what follows I will be interested exclusively in the former of these alternatives and I will focus my attention on the *logic of knowledge*.

Working with an epistemological interpretation of modal logic, it is worth specifying who is the subject of the knowledge we are speaking about. If we read $\Box p$ simply as "it is known that p", the meaning of this operator remains not clear enough. What does it mean, in fact, that something is known? Does it mean that someone knows it? Or does it mean that everyone knows it? Therefore, in order to be as clear as possible, we should adopt a more intuitive terminology and make explicit the fact that we are working with a *propositional notion of knowledge* and within a logic of *individual agents*. The box operator will be substituted by a K (for "knowledge"), followed by a letter that indicates who is the agent that knows the considered proposition. Modal formulas will look, thus, like $K_a p$ and $K_b p$ and they will be read as "the agent a knows that p" and "the agent b knows that p". In what follows, we will be interested in formal systems with only one agent, but it is important to keep in mind that we can introduce many K-operators, in order to map the knowledge of more than one subject[1].

Let us now move, after these introductory remarks, to give a precise definition of the *syntax* of the propositional modal logic for knowledge. We proceed extending the alphabet of classical propositional logic with a knowledge operator K_a.

Definition 2.1 (Alphabet of Propositional Modal Logic for Knowledge)**.** An alphabet for propositional modal logic for knowledge is defined as the union of the following disjointed sets:

- A denumerable set of *atomic propositional variables* $\mathcal{P} = \{p_0, p_1, ...\}$.
- The set of the *logical connectives* $C = \{\neg, \wedge, \rightarrow\}$.
- The set of the *knowledge operator* $O = \{K_a\}$.
- The set of *auxiliary symbols* $\mathcal{A} = \{(,)\}$.

Given the alphabet, it is possible to define inductively the set of the formulas of the logic of knowledge.

Definition 2.2 (Formulas of Propositional Modal Logic of Knowledge)**.** The formulas of the modal logic of knowledge are given by the following definition by induction:

1. If φ is an atomic propositional variable, then φ is a formula.
2. If φ is a formula, then also its negation $\neg\varphi$ is a formula.
3. If φ and χ are formulas, then also their conjunction $(\varphi \wedge \chi)$ is a formula.
4. If φ and χ are formulas, then also the conditional $(\varphi \rightarrow \chi)$ is a formula.
5. If φ is a formula, then also $K_a\varphi$ is a formula.

[1] For the introduction of multiple agents see both Hendricks and Symons, (2015, pp. 9-11) and Holliday, (forthcoming, pp. 5-7).

6. Nothing else is a formula.

The *semantics* of the logic of knowledge is provided by a Kripke-structure, which is the standard way to interpret modal languages.

Definition 2.3 (*Kripke-structures*)**.** Given a propositional modal logic of knowledge, a *Kripke-structure* $\mathcal{M}$ is a triple $\langle W, R_a, V \rangle$, where:

1. W is a non-empty set. Intuitively, W is a set of "possible worlds" or "possible scenarios".

2. R_a is a binary relation over W, i.e. a subset of $W \times W$. Intuitively, we read vR_aw as "the possible world w is epistemically accessible from the possible world v by the agent *a*".

3. V is a function that assigns to every atomic propositional formula a subset of W. Intuitively, V specifies in which possible worlds each atomic formula is true.

Given the Kripke-structures, we can define the notion of truth in a world:

Definition 2.4 (*Truth in a world*)**.** Given a propositional modal logic for knowledge, a Kripke-structure $\mathcal{M}$ and a world w, the notion $\mathcal{M} \vDash_w \varphi$ of being true in a world is defined as follows:

1. when φ is atomic, then $\mathcal{M} \vDash_w \varphi$ iff $w \in V(\varphi)$;

2. when φ has the form $\neg\chi$, then $\mathcal{M} \vDash_w \varphi$ iff $\mathcal{M} \nvDash_w \chi$;

3. when φ has the form $(\chi \wedge \psi)$, then $\mathcal{M} \vDash_w \varphi$ iff $\mathcal{M} \vDash_w \chi$ and $\mathcal{M} \vDash_w \psi$;

4. when φ has the form $(\chi \rightarrow \psi)$, then $\mathcal{M} \vDash_w \varphi$ iff $\mathcal{M} \nvDash_w \chi$ or $\mathcal{M} \vDash_w \psi$;

5. when φ has the form $K_a\chi$, then $\mathcal{M} \vDash_w \varphi$ iff for every possible world v such that wR_av, $\mathcal{M} \vDash_v \chi$.

The definition of truth in a world allows us to define two further important notions. We say that a formula φ is *true in a model* $\mathcal{M}$ if and only if it is true in every world $w \in W$ of the Kripke-structure $\mathcal{M}$. We say that a formula φ is a *valid formula* if and only if it is true in every world $w \in W$ of every Kripke-structure $\mathcal{M}$.

What we have described so far is the minimal system **K** of modal logic, with the only peculiarity that the informal reading that we have assumed for the modal operator is "the agent *a* knows that...". Nevertheless, it is clear that to obtain a logic of knowledge this is not enough. What one needs, rather, is to specify the formal properties that are typical of knowledge and to represent them in the logic. Putting specific restrictions over the accessibility relation R_a, it is possible to obtain many modal logics stronger than **K**, where more principles are valid formula. The problem is that it is not sufficiently clear which modal system photographs in the correct way the formal properties of knowledge. Since the purpose of this article is to examine which logic can be accepted by epistemologists with different philosophical orientations, we will extend **K** only with those principles which are generally taken for granted in the epistemological debate. Therefore, the only restriction that we want impose to our logical system is that it has to satisfy the following principle:

(T) $K_a\varphi \to \varphi$

What (T) says is that, if one knows a proposition, then this very same proposition must be true. This does not only follow from any analysis of knowledge as true belief plus something, but it also seems to be a valid minimal description of the meaning of knowledge. Indeed, if one says that he/she knows that p but it is not the case that p, it seems reasonable to conclude that he/she *does not* know that p, but rather only *believes* that p[2].

If we want that the principle (T) holds in the logical framework that we are considering, we have to put a restriction on the accessibility relation R_a. More precisely, as we prove with the following theorem, we have to restrict our attention to those Kripke-structures where the accessibility relation is reflexive. The modal logic that we obtain when we work only with reflexive accessibility relations is called **T**.

Theorem 2.1. *Given the language of propositional modal logic and its Kripke-structure $\mathcal{M} = \langle W, R_a, V\rangle$, the formula (T) $K_a\varphi \to \varphi$ is a valid formula if and only if the accessibility relation R_a is reflexive.*

Proof: Assuming that the accessibility relations R_a in $\mathcal{M}$ is reflexive, then given any possible world $w \in W$ we have that wR_aw. Therefore, since $\mathcal{M} \vDash_w K_a\varphi$ holds, then $\mathcal{M} \vDash_v \varphi$ holds in every world v such that v is accessible from w. But for reflexivity we have that w is accessible from itself and, therefore, that $\mathcal{M} \vDash_w \varphi$. *Vice versa*, assuming that $K_a\varphi \to \varphi$ is a valid formula then, for every Kripke-structure $\mathcal{M}$ and every world w in it $\mathcal{M} \vDash_w K_a\varphi \to \varphi$. Given the semantics of the conditional, this amounts to say that it is not the case that $\mathcal{M} \vDash_w K_a\varphi$ and $\mathcal{M} \nvDash_w \varphi$. But, if R_a was not reflexive, we could construct a Kripke-structure such as $\mathcal{N} = \langle W, R_a, V\rangle$, with $W = \{v, w\}$ and $R_a = \{\langle w, v\rangle\}$. In $\mathcal{N}$ we have that, if $v \in V(\varphi)$ but $w \notin V(\varphi)$, then $\mathcal{N} \vDash_w K_a\varphi$ but $\mathcal{N} \nvDash_w \varphi$, contradicting our claim that $K_a\varphi \to \varphi$ is a valid formula. Therefore, R_a must be reflexive. ■

3 The principle of epistemic closure and its problems

In the previous part of this article I have introduced the modal logic **T**, in order to represent some minimal formal properties of knowledge. Moving a step further,

[2] This aspect is famously stressed by Wittgenstein, (1969).

it is now possible to prove an interesting result, which says that the principle of epistemic closure is a valid formula in **T**. Firstly, let us clarify what we mean with the name of "principle of epistemic closure".

(CP) If an agent knows that φ and he/she knows that if φ then χ, then he/she also knows that χ.

It is straightforward to translate this thesis into the language of the logic of knowledge. We thus obtain the following formal version of the closure principle:

(FCP) $(K_a\varphi \land K_a(\varphi \to \chi)) \to K_a\chi$

We can now prove the following theorem:

Theorem 3.1. *Given the logic of knowledge* **T***, the formal closure principle (FCP) is a valid formula.*

Proof: We reason for absurd. If (FCP) was not a valid formula, there would be a world w of a Kripke-structure $\mathcal{M} = \langle W, R_a, V\rangle$, where (FCP) does not hold. Given the semantics of the conditional, this means that $\mathcal{M} \vDash_w K_a\varphi \land K_a(\varphi \to \chi)$ but $\mathcal{M} \nvDash_w K_a\chi$. Given $\mathcal{M} \vDash_w K_a\varphi$, we have that in every world accessible from w, φ holds. Given $\mathcal{M} \vDash_w K_a(\varphi \to \chi)$, we have that in every world accessible from w, $\varphi \to \chi$ holds. Moreover, since $\mathcal{M} \nvDash_w K_a\chi$, there is at least one world v such that wR_av where $\mathcal{M} \nvDash_v \chi$. But in this same world v we have that $\mathcal{M} \vDash_v \varphi$ and $\mathcal{M} \vDash_v \varphi \to \chi$ hold too, from which it follows that $\mathcal{M} \vDash_v \chi$. Therefore, we obtain the contradiction that $\mathcal{M} \vDash_v \chi$ and $\mathcal{M} \nvDash_v \chi$. ■

If our concerns are mainly epistemological this result has a particular relevance. In fact, what we have proved is that even if we work with a weak modal system, the principle of epistemic closure will hold in it[3]. Therefore, if we have some reason to refuse the principle of epistemic closure, then we can not adopt the formal logic **T** anymore, for it describes knowledge in a way which is inconsistent with our theory. In particular Dretske (1970) offers at least two possible reasons to refuse the closure principle[4]. In the rest of this part I will present both of them, but I will not try to set the question about their validity. Indeed, I only want to show that it might be reasonable for an epistemologist to reject the closure principle. In fact, given the possibility that (FCP) is not acceptable, we have to look for a modal logic for knowledge weaker than the standard one described

[3] Notice, moreover, that in the proof of the theorem 3.1. we did not make any use of the fact that the accessibility relation between worlds is reflexive. Therefore, our proof is valid also for the basic modal logic **K**.

[4] Luper, (2016) synthesizes a wide range of arguments against the closure principle, often originally raised by Dretske and Nozick. However, even if Luper's reconstruction is clear, I do not agree with his presentation of the arguments from the "analysis of knowledge". In fact, the theories of knowledge supported by Dretske and Nozick are *explanations* of why the closure principle fails and not *reasons* to refuse it. Luper commits, therefore, a sort of inversion of the right order of explanation.

by the Kripke-structures. Our purpose, in fact, is not to take part in the epistemological debate and to identify the modal logic which better describes knowledge but, rather, it is to find a minimal logical framework which can be accepted by epistemologists of different currents.

A first critique to the principle of epistemic closure is linked to skepticism. In fact, one general way to reconstruct the argument presented by the skeptic is with the following argument:

(1) I do not know that I am not a brain in a vat
(2) If I do not know that I am not a brain in a vat, then I do not know that I have hands.

---

(3) I do not know that I have hands $\therefore$

The premiss (2) of this argument is a consequence of an instance of (CP). If I know that I have hands and I know that if I have hands I am not a brain in a vat, then I know that I am not a brain in a vat. Therefore, if I do not know that I am not a brain in a vat, then either I do not know that I have hands, or I do not know that if I have hands I am not a brain in a vat. However, since I know that if I have hands I am not a brain in a vat, we can exclude the second disjunct and obtain (2): if I do not know that I am not a brain in a vat, then I do not know that I have hands[5].

If skepticism is expressed in the form of the syllogism presented above, there are two main strategies to criticize it. Either one denies the premiss (1), either one denies the premiss (2), namely the closure principle. The first horn was chosen by Moore (1939), who reversed the skeptic's argument in its contraposed version[6].

(1) I do know that I have hands
(2) If I do not know that I am not a brain in a vat, then I do not know that I have hands.

---

(3) I do know that I am not a brain a vat $\therefore$

[5] It is worth underlining that, in order to obtain (2) from (CP), we have to take for granted that we know that if we have hands we are not a brain in a vat. Although this might seem trivial, there are two problematic aspects which deserve some further reflections. On the one hand, one may think that it is much more reasonable to deny the premiss of the argument from (CP) to (2), namely to assert that we do not know that if we have hands then we are not a brain in a vat, rather than to accept the conclusion it leads to, i.e. that we do not know that we have hands. On the other hand, there might be a skeptical scenario that we do not know, or a person who never thought about brains in a vat. But if one has never thought about a skeptical scenario, it does not seem plausible to say that he/she knows that if he/she has hands, then he/she is not in the considered skeptical scenario.

[6] For historical's sake, let me remark that Moore did not deal with the brain in a vat hypothesis in his original article of 1939, but he rather considered more traditional skeptical scenarios.

However, this solution implies that we do actually know that we are not brains in a vat, which is a conclusion that many might find excessively strong. Therefore, if we want to remain faithful both to the intuition that we do know that we have hands, both to the intuition that we do not know that we are not brains in a vat, we have to abandon the closure principle. Notice that this is not an argument against skepticism. If we want to criticize skepticism *because* the closure principle does not hold we need independent arguments against (CP). On the contrary, this is an argument against the closure principle, *because* skepticism does not hold. So, what this argument needs are independent reasons to refuse skepticism.

However, Dretske criticizes the principle of epistemic closure also in a second more explicit way, bringing some counterexamples to it. Perhaps the most famous one is the so-called "zebra case". Imagine that you are in a zoo with your nephew. While you are walking around, he asks you if you know what is the animal you are looking at. You observe it, you notice that it looks exactly how you expect a zebra should look like, and you also find a sign with "zebra" written on it. Without any further doubt you would reply to your nephew's question something like: "Sure! It is a zebra". Thus, you do know that the animal you are observing is a zebra. But do you know that it is not a disguised mule? Indeed, it might be a mule so well depicted by the zoo-officers to look exactly like a zebra, maybe in order to attract more visitors.

Examples like this present a sort of strange situation. On the one hand, we have a plenty of reasons to believe that the animal we are observing is a zebra. On the other hand, we do not know that it is not a disguised mule. Moreover, we are also completely aware that mules and zebras are different animals. Therefore:

(i) we know that the animal we are looking at is a zebra;

(ii) we know that if the animal we are looking at is a zebra, then it is not a disguised mule;

(iii) we do not know that the animal we are looking at is not a disguised mule.

Clearly, (i), (ii) and (iii) taken together are an instance of failure of the closure principle.

Together, these two arguments show that the principle of epistemic closure is not so obvious and trivial as one might believe at first sight. A closer examination of it shows both that it has skeptical consequences and that it does not always fit our intuitions in concrete examples. Therefore, if we want to find a propositional modal logic which describes some minimal properties of knowledge generally accepted by epistemologists we have to weaken in some way the logic of knowledge that we have previously presented.

4 The impossible worlds and the Rantala-semantics

In the context of the logical literature, an alternative to the standard Kripke-semantics has been provided in order to account for the problem of logical omniscience. In fact, one further consequence of adopting a modal logic like **K** or stronger is that any agent knows every classical tautology. In fact, since classical tautologies are valid in every possible world, the agent always knows them, for they are trivially true in all the worlds which the agent has access to. Although it is important to keep distinct the problem of the epistemological closure principle from the one of the logical omniscience, we can try to apply the logical system used to answer to the latter of these problems also to respond to the former one[7].

Given the syntax of modal logic that we have already defined, we can introduce a slightly different semantics, namely a Rantala–semantics[8].

Definition 4.1 (*Rantala-structures*)**.** Given a propositional modal logic of knowledge, a *Rantala-structure* $\mathcal{R}$ is a quadruple $\langle W, W', R_a, V\rangle$, where:

1. W is a non-empty set. Intuitively, W is a set of "possible worlds" or "possible scenarios".

2. W' is a set. Intuitively, W' is a set of "impossible worlds" or "impossible scenarios".

3. R_a is a binary relation over $W \cup W'$, i.e. a subset of $(W \cup W') \times (W \cup W')$. Intuitively, we read vR_aw as "the possible or impossible world w is epistemically accessible from the possible or impossible world v by the agent a".

4. V is a function that assigns to every atomic propositional formula a subset of $W \cup W'$ and to every formula a subset of W'. Intuitively, V specifies in which possible or impossible worlds each atomic formula is true, and in which impossible worlds each formula is true.

As one can immediately notice, the difference between the Kripke and the Rantala structures relies on the introduction of a set of impossible worlds. To see how they affect the interpretation of every formula, we shall reformulate also the notion of truth in a model.

Definition 4.2 (*Truth in a world*)**.** Given a propositional modal logic for knowledge, a Rantala-Structure $\mathcal{R}$ and a world w, the notion $\mathcal{R} \vDash_w \varphi$ of being true in a world is defined as follows:

[7] On the difference between the problem of logical omniscience and the one of epistemic closure see Holliday, (forthcoming, pp. 8-10).

[8] The name of Rantala-semantics comes from the Finnish logician Veikko Rantala. Here I follow the presentation of its semantics given by Wansing, (1990), who also provides an interesting comparison between the Rantala-semantics and other methods to solve the paradox of logical omniscience.

1. If $w \in W'$, namely if w is an impossible world, then $\mathcal{R} \vDash_w \varphi$ iff $w \in V(\varphi)$;

2. If $w \in W$, namely if w is a possible world, then:

   (a) when φ is atomic, then $\mathcal{R} \vDash_w \varphi$ iff $w \in V(\varphi)$;

   (b) when φ has the form $\neg\chi$, then $\mathcal{R} \vDash_w \varphi$ iff $\mathcal{R} \nvDash_w \chi$;

   (c) when φ has the form $(\chi \wedge \psi)$, then $\mathcal{R} \vDash_w \varphi$ iff $\mathcal{R} \vDash_w \chi$ and $\mathcal{R} \vDash_w \psi$;

   (d) when φ has the form $(\chi \to \psi)$, then $\mathcal{R} \vDash_w \varphi$ iff $\mathcal{R} \nvDash_w \chi$ or $\mathcal{R} \vDash_w \psi$;

   (e) when φ has the form $K_a\chi$, then $\mathcal{R} \vDash_w \varphi$ iff for every possible or impossible world v such that wR_av, $\mathcal{R} \vDash_v \chi$.

It is now possible to clarify which is the role that the impossible worlds play in the new structure now defined. A first notable aspect is that, while in regards of the possible worlds the notion of truth in a world is defined inductively, the truth-value of every formula in an impossible world is directly specified by the assignment V. In an impossible world we might have that a disjunction is true even if its two disjuncts are both false, or that even if two formulas are true their conjunction is false, and so on. The distinguished aspect of this structure is that the anomalous behaviour of impossible worlds has some consequences on the evaluation of formulas in "normal" possible worlds. In fact, in order for a modal formula like K_ap to be true in a possible world w, the formula p has to be true in every world v, both possible and impossible, such that wR_av.

The notion of valid formula has now to be defined for the new Rantala-semantics: we say that a formula φ is a *valid formula* if and only if it is true in every possible world of every Rantala-structure. Given this new definition and thanks to the introduction of the impossible worlds, we can show that the principle of epistemic closure (FCP) is not a valid formula anymore. In fact, even if $\mathcal{R} \vDash_w K_a\varphi$ and $\mathcal{R} \vDash_w K_a(\varphi \to \chi)$, it is still possible that $\mathcal{R} \nvDash_w K_a\chi$, since there might be an impossible world i such that wR_ai where $i \in V(\varphi)$ and $i \in V(\varphi \to \chi)$ but $i \notin V(\chi)$.

Moreover, notice that the introduction of impossible worlds does not imply that "everything goes". We can, as we have already done for **K**, propose a strengthening of this logical framework in order to meet at least the essential properties of the knowledge operator. Exactly as we have argued in the first part of this article, the minimal requirement for a logic of knowledge seems to be that if we know a proposition, then this very proposition is true. Again, if we impose that the accessibility relation is reflexive, then we obtain a logic where the formula (T) $K_a\varphi \to \varphi$ is a valid formula. In this way we can define the new logic **T'**, obtained by considering only those Rantala-structures where the accessibility relation between worlds is reflexive.

5 An evaluation of the Rantala-semantics strategy

In this last part I shall draw some consequences from the previous analysis and try to evaluate if the Rantala-semantics that we have defined provides a minimal logical framework to describe the formal properties of knowledge. Firstly, I argue that it is possible to identify two reasons to believe that the Rantala-semantics actually describes a valid minimal logic of knowledge. Then I will consider two objections. While one will result to be only an apparent critique to the Rantala-semantics strategy, the second one will identify a true limit of it.

(i) A first observation is that the logic **T'** that we have defined actually provides the minimal logical framework for knowledge which we were looking for. On the one hand, the principle (T) $K_a\varphi \rightarrow \varphi$ results to be a valid formula in this system: working in **T'** we can represent the fact that if an agent knows a proposition, then that proposition is true. On the other hand, the logic **T'** does not force us to accept the closure principle, since (FCP) is not a valid formula in it. Therefore, epistemologists with different theories about knowledge can all accept the modal system **T'** as a minimal framework, which reflects only those properties of knowledge which are unanimously recognized.

(ii) Moreover, the Rantala-semantics is sufficiently flexible to provide not only a minimal common framework, but also a basis suitable for further developments. Given the minimal logic **T'**, it is possible to obtain systems with new axioms or inference rules imposing new conditions on the accessibility relation R_a or on the evaluation function V[9]. In this way, the Rantala-semantics can be used also to represent more complex theories of knowledge, in which more principles hold and should be treated as valid formulas. Epistemologists of different philosophical orientations will thus share the common framework given by **T'**, and they will also be able to describe more complex and rich systems without the need of describing a new and different semantics. Even if **T'** is a quite general and minimal system, we can start from it and obtain step by step new and stronger logics, which will formalize richer and more complex accounts of knowledge.

(iii) However, one aspect of the Rantala-semantics that some philosophers may find problematic is the fact that it makes use of impossible worlds. In fact, even if we accept to work with the framework of possible worlds of the Kripke-structures, the introduction of impossible worlds poses some new problems. Indeed, although possible worlds represent sets and combinations of facts and events that are not actual, they are still consistent with the laws of classical logic. Differently, it is not straightforward to account for worlds where the most evident logical contradictions may hold. In an impossible world both a proposition and its negation might be true, two disjuncts can be true and the entire disjunction

[9] Compare with Wansing, (1990), who also presents some examples of restriction.

false, and so on. Nevertheless, even if impossible worlds surely present paradoxical features, I think that this problem is only apparent.

Firstly, as Nolan (2013, p. 367–370) underlines, almost every metaphysical theory about the possible worlds can be extended in order to account also for the impossible ones. The only theory which has some problems while explaining the nature of impossible worlds is modal realism, which regards possible worlds as entities really existing. However, there are also some attempts to extend the modal realist perspective in order to describe impossible worlds[10]. Moreover, one may also decide to follow an alternative direction and to consider the useful theoretical role of the impossible worlds a valid reason to reject modal realism and to defend another metaphysical perspective also in regards of the "normal" possible worlds.

Furthermore, it is not obvious at all that the introduction of impossible worlds in epistemic logic forces us to take an explicit position about their metaphysical nature[11]. In fact, the specific philosophical problems that a modal logic raises are linked to the informal interpretation that we decide to give of its operators. For instance, if we read the box symbol as representing necessity, then we have to clarify what does it mean that a proposition is necessary in a world w if and only if it is true in every possible world which is accessible from w. An analysis of the nature of possible world is essential, in this case, in order to make sense of the metaphysical interpretation of the system of modal logic that we are considering. However, if the reading that we are adopting is epistemic, we do not need to take such a metaphysical attitude. As we have already said defining the Kripke-structures, the label of possible world can be substituted without any problem with the one of "scenario". Indeed, the possible and impossible worlds are only the combinations of facts and events that an agent may find plausible descriptions of the reality or not. The informal epistemological reading of the knowledge operator does not call for any metaphysical interpretation. The fact that an agent knows a proposition if and only if that proposition is true in every world to which he/she has access only means that that proposition is part of all the descriptions that he/she considers as possibly valid representations of the reality.

(iv) Ultimately, despite its many virtues, I think that it is possible to identify a proper limit of the Rantala-semantics strategy. Let us distinguish two different aspects: the failure of the closure principle itself and the explanation of the fact that it does not hold. Depending on what we ask to an epistemic logic, we might then give different evaluations to the Rantala-semantics strategy. On the one hand, as I have already pointed out, the modal logic **T'** offers a formal sys-

[10] Compare with Nolan, (2013, p. 369).

[11] Wansing, (1990, p. 536) takes an even stronger position, saying that the question itself about the nature of the impossible worlds is "unsatisfactory".

tem where the closure principle of knowledge is not a valid formula. If we adopt **T'**, indeed, we are able to represent many formal properties of knowledge and to potentially adjust the system – working on the accessibility relation and the evaluation function – to meet the characteristics of different epistemological theories. On the other hand, the Rantala-semantics does not provide an explanation of why the closure principle fails. Or, even worse, one may argue that it actually gives a *wrong* explanation of this fact. Indeed, the "cause" that determines the failure of (FCP) in the Rantala-semantics is the introduction of the impossible worlds. If we try to interpret this formal aspect from an epistemological perspective, the result is that the epistemic closure principle does not hold because the agent consider as plausible descriptions of the reality also scenarios where the laws of logic do not hold. However, the problem is that this is not the explanation that the epistemologists who refuse closure – notably Dretske and Nozick – have provided. Therefore, even if it offers a framework that can be accepted also by the epistemologists who do not accept the closure principle, the Rantala-semantics do not reflect in any way their intuitions about why this principle does not hold[12].

Finally, trying to sum up the considerations developed in this last part, it is possible to sketch an evaluation of the Rantala–semantics strategy. The result that we obtained can be regarded as twofold and it depends on what we ask to an epistemic logic. If we want a strong characterisation of a formal system, such that it reflects all the theoretical features of an epistemological theory, then the Rantala–semantics strategy does not seem to be the right way to account for the problems presented by the closure principle. Still, a more modest attitude is also possible. In fact, we can demand to a formal system only to verify as valid those principles – and only those – which an epistemological theory regards as the formal properties of knowledge. In this light, even if it does not provide any heuristic insight about the failure of (FCP), the Rantala-semantics is an interesting common framework for different epistemological perspectives, which can also be refined and strengthened in further ways.

[12] An interesting contribution on this topic is Holliday, (2015), who directly formalizes the epistemological theories proposed by Dretske and Nozick. Notice, however, that although in this way a formal system gains in heuristic power, it also loses the generality that makes it acceptable by epistemologists with different ideas.

References

Dretske, Fred I. (1970). "Epistemic Operators". In: *The Journal of Philosophy* 67.24, pp. 1007–1023.

Hendricks, Vincent and John Symons (2015). "Epistemic Logic". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Holliday, Wesley H. (2015). "Epistemic Closure and Epistemic Logic I: Relevant Alternatives and Subjunctivism". In: *Journal of Philosophical Logic* 44.1, pp. 1–62.

— (forthcoming). "Epistemic Logic and Epistemology". In: *Handbook of Formal Philosophy*. Ed. by Sven Ove Hansson and Vincent F. Hendricks. Springer.

Luper, Steven (2016). "Epistemic Closure". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Moore, George E. (1939). "Proof of an External World". In: *Proceedings of the British Academy* 25, pp. 273–300.

Nolan, Daniel P. (2013). "Impossible Worlds". In: *Philosophy Compass* 8.4, pp. 360–372.

Wansing, Heinrich (1990). "A general possible worlds framework for reasoning about knowledge and belief". In: *Studia Logica* 49.4, pp. 523–539.

Wittgenstein, Ludwig (1969). *Über Gewißheit*. Ed. by Elizabeth Anscombe and Georg Henrik von Wright. Frankfurt am Main (DEU): Suhrkamp.

Sponsored since 2011 by the
Italian Society for Analytic Philosophy

ISSN 2037-4445
http://www.rifanalitica.it

A New Model for the Liar

Luca Castaldo[*]

Abstract. A new model for $\mathscr{L}_{pa}^{t}$ (the language of arithmetic enhanced by the unary truth predicate T) is presented, which extends Kripke's minimal fixed point. The latter, it will be argued, does not adequately model the truth predicate, since no difference between Liars and Truth-tellers can be made. The new model, which contains an extension of Kripke's interpretation of T along with a new 4-valued logic, overcomes this inadequacy. The gist of my proposal is that 'paradoxical' ought to be treated as a truth value: Liar sentences, unlike Truth-teller sentences, do not simply *lack* a truth value. They do posses one: they are *paradoxical*.

Keywords. Formal theories of truth, Liar paradox, 4-valued logic, Kripke.

[*] I wish to thank Professor Karl-Georg Niebergall for valuable comments and suggestions.

Author. Luca Castaldo, castaldx@gmail.com.

Received. 7 September 2016. **Accepted**. 29 October 2016.

Introduction

It seems that

> **P** Every sentence is *either* true *or* untrue[1].

But, if so, what about the Liar sentence below?

> **$** The sentence marked with a dollar is untrue.

$ contradicts **P**, for it is true *if, and only if,* it is untrue. To (dis)solve the problem, Kripke (1975) proposes to reject **P** and, exploiting the 3-valued logic called *Strong Kleene*, constructs a partial model for the truth predicate, where sentences like **$** are 'undefined', i.e. they *lack* a truth value.

Within Kripke's model, however, also the so-called Truth-teller

> € The sentence marked with a euro is true.

lacks a truth value. Yet, **$** and **€** are, admittedly, very different: the latter can *consistently* be declared true or untrue; the former cannot. An adequate model for the truth predicate ought to account for their diversity.

The purpose of this paper is to put forward a new response to the Liar paradox, which extends and improves the work done by Saul Kripke in his seminal *Outline of a Theory of Truth*.

The plan is as follows: after technical preliminaries in § 1 (including the construction of the formal Liar sentence), I go on in § 2 to present a new model for the truth predicate along with a new 4-valued logic, thereby proposing the new response to the Liar paradox. The final section 3 examines the properties of the model, proving what I shall call 'metalinguistic T-Schema'.

A last remark before I begin: In what follows I assume the reader is familiar with (i) *Peano arithmetic*, (ii) the *arithmetization of syntax*, and (iii) Kripke's *Outline of a Theory of Truth*[2].

[1] The *either ... or* is to be read here as an exclusive disjunction.

[2] There is an extensive literature on Kripke's *Outline*. A more philosophical and informal introduction is offered by Burgess, (2011). For more information on the mathematical aspects of Kripke's construction see, for example, Fitting, (1986) and McGee, (1991, §§4-5). The axiomatic theory known as Kripke-Feferman (**KF**) was first given by Reinhardt, (1986) and Feferman, (1991). Feferman, (1991) also determines its proof-theoretic strength. Cantini, (1989) gives a more direct proof-theoretic analysis of **KF** and some of its subsystems. In **KF**, the *partial* notion of truth advanced by Kripke is axiomatised in *classical* logic. Therefore, outer logic (what is provable) and inner logic (what is provably true) of that system differs substantially. Halbach and Horsten, (2006) (see also Horsten, 2011, §9.5) have proposed an interesting axiomatisation in partial logic, creating a system, called "partial Kripke-Feferman" (**PKF**), within which the two logics coincide. In that system, gaps but no gluts are admitted. Halbach, (2014, §16) proposes a system that admits both. For critical discussions of Kripke's position see, among others, Gupta, (1982, pp. 30-37) and Field, (2008, §3).

1 The Formalised Liar

1.1 Technical Preliminaries

The object language of this work will be the language of Peano arithmetic (**PA**) extended by the unary truth predicate T. I shall call the language of **PA**, without T, $\mathscr{L}_{pa}$; the extended language will be called $\mathscr{L}^{t}_{pa}$[3]. As "official" logical vocabulary, I shall use the existential quantifier $\exists$, the negation and disjunction symbols $\neg$, $\vee$, and the identity symbol $\doteq$. As usual, however, abbreviations will be used. A standard *Gödel numbering* of $\mathscr{L}^{t}_{pa}$-expressions will be assumed throughout the work, without going into details[4]. The Gödel number (or code) of a formula φ is $gn(\varphi)$, and $\ulcorner\varphi\urcorner$ is the numeral of $gn(\varphi)$. I shall distinguish between natural numbers and $\mathscr{L}^{t}_{pa}$-numerals exploiting boldfaced characters: the natural numbers are written "$0, 1, 2, \dots, n$" (not boldfaced) and the $\mathscr{L}^{t}_{pa}$-numerals "$\mathbf{0}, \mathbf{1}, \mathbf{2}, \dots, \mathbf{n}$" (boldfaced), where "$\mathbf{1}, \mathbf{2}, \mathbf{3}\dots$" abbreviates "$\mathbf{0}', \mathbf{0}'', \mathbf{0}'''\dots$". Formulae with one free variable are indicated by $\varphi(v_i)$; $\varphi(t)$ denotes $\varphi[t/v_i]$, i.e. the result of substituting t for v_i in φ. I write $\varphi \equiv \psi$ to indicate that φ and ψ are names of the same formula.

$\langle \mathcal{M}, (E_\infty, A_\infty)\rangle$ is Kripke's minimal fixed point (henceforth: Mfp), and '$\langle \mathcal{M}, (E_\infty, A_\infty)\rangle \models_{\text{sk}} \varphi$' means that φ is true in Mfp, according to the Strong Kleene. Furthermore, I shall make use of the following metalinguistic symbols:

- $\neg\!\!\neg$ for "non …".
- $\vee\!\!\!\vee$ for "… or …".
- $\wedge\!\!\!\wedge$ for "… and …".
- $\Rightarrow$ for "if …, then …".
- $\Leftrightarrow$ for "… if, and only if, …".
- $\exists\!\!\!\exists$ for "there is …".
- $\forall\!\!\!\forall$ for "for all …".

1.2 $\lambda \leftrightarrow \neg T\ulcorner\lambda\urcorner$

The *Diagonal Lemma*[5] is, as McGee, (1991, p. 24) put it, "a cornerstone of modern logic". He even adds that "most of the results of [*Truth, Vagueness, and Paradox*] can be regarded as corollaries to this basic result". In this section I shall exploit the typical diagonal construction, in order to obtain the formalised liar antinomy.

[3] Notice that we are just extending the language of **PA**, not the theory, i.e. we are not adding axioms for T, creating a new theory, say $\mathbf{PA}^{\mathbf{T}}$. In addition, we can impose a restriction on the induction schema to $\mathscr{L}_{pa}$-formulae, i.e., an instance of

$$(\varphi(\mathbf{0}) \wedge \forall v_i(\varphi(v_i) \rightarrow \varphi(v_i'))) \rightarrow \forall v_i(\varphi(v_i))$$

is an axiom, only if T does not occur in φ.

[4] See, for instance, Boolos, Burgess, and Jeffrey, (2007) and Smith, (2013).

[5] Or *Fixed Point Lemma*, or *Self-Referential Lemma*.

Before I begin, the concept of *diagonalization* of a formula must be introduced:

> The *diagonalization* of φ is the expression $\exists v_0(v_0 \doteq \ulcorner\varphi\urcorner \wedge \varphi)$.

Even if this notion makes sense for arbitrary expressions, it is of most interest in the case of a formula $\varphi(v_0)$ with just one variable v_0 free. Since an expression of the form $\varphi(t)$ is equivalent to $\exists v_0(v_0 \doteq t \wedge \varphi(v_0))$, the diagonalization of $\varphi(v_0)$ is equivalent to $\varphi(\ulcorner\varphi\urcorner)$. That is: the diagonalization of a formula $\varphi(v_0)$ is true (in the standard interpretation) if, and only if, it is satisfied by its own code.

There is also a recursive function $diag$ that, when applied to the Gödel number of a formula, yields the Gödel number of its diagonalization. That is to say: if the code of a formula φ is n and the code of its diagonalization is m, then $diag(n) = m$. A more formal definition is:

$$diag(n) = gn[\exists v_0(v_0 \doteq\ ] \star num[n] \star gn[\ \wedge\ ] \star n \star gn[)],$$

where $\star$ and num represent, respectively, the concatenation and the numeral functions, both recursive[6].

Lemma 1.1. (The Formalised Liar) There is a $\mathscr{L}^t_{pa}$-sentence λ, such that

$$\mathbf{PA} \vdash \lambda \leftrightarrow \neg T\ulcorner\lambda\urcorner$$

Proof. Since **PA** represents every primitive recursive function, $diag$ is representable in **PA**. Let $\mathbf{Diag}(v_0, v_1)$ be a formula representing $diag$, so that for any a and b, if $diag(a) = b$, then

$$\mathbf{PA} \vdash \forall v_1(\mathbf{Diag}(\mathbf{a}, v_1) \leftrightarrow v_1 \doteq \mathbf{b}) \tag{1}$$

Diag is a complex $\mathscr{L}_{pa}$-formula, *not* containing the new predicate T.

Let now $\beta(v_0)$ be the formula

$$\exists v_1(\mathbf{Diag}(v_0, v_1) \wedge \neg T(v_1)) \tag{$\beta(v_0)$}$$

Intuitively, $\beta(v_0)$ says that the diagonalization of *a* formula is not true, without yet saying *which* formula. Let's now consider the diagonalization of $\beta(v_0)$, and let's call it λ:

$$\exists v_0(v_0 \doteq \ulcorner\beta\urcorner \wedge \beta(v_0)) \tag{λ}$$

[6] The concatenation function $\star$ is such that, if s and t are the codes of two expressions, then $s \star t$ is the code of the first expression followed by the second. The numeral function num maps each n to the code of the numeral $\mathbf{n}$. The function $diag$ could have been defined more precisely by first showing that also the logical operations of conjunction and existential quantification are recursive. For more information see Boolos, Burgess, and Jeffrey, (2007, p. 221, §15).

In other symbols, λ is

$$\exists v_0\big(v_0 \doteq \ulcorner\beta\urcorner \wedge \exists v_1(\mathbf{Diag}(v_0, v_1) \wedge \neg T(v_1))\big)$$

This is logically equivalent to $\beta(\ulcorner\beta\urcorner)$, i.e. the result of substituting $\ulcorner\beta\urcorner$ for v_0 in $\beta(v_0)$:

$$\exists v_1(\mathbf{Diag}(\ulcorner\beta\urcorner, v_1) \wedge \neg T(v_1)) \tag{2}$$

Reading (2) in English, we get something like: "there is a number that has two properties: first, it is the code of the diagonalization of $\beta(v_0)$; second, it is not element of the extension of T". Or, more intuitively: "the diagonalization of β is not true". Interesting enough, the diagonalization of β is precisely λ. Accordingly, λ is logically equivalent to a sentence that says that λ is not true.

We have thus far constructed, within the formal language $\mathscr{L}^t_{pa}$, a sentence saying of itself that it is not true[7]. The next step consists in proving, within **PA**, something about this sentence. Since λ is logically equivalent to (2), we have:

$$\mathbf{PA} \vdash \lambda \leftrightarrow \exists v_1(\mathbf{Diag}(\ulcorner\beta\urcorner, v_1) \wedge \neg T(v_1)) \tag{3}$$

We do not know, whether λ is a theorem of **PA**. We do know, however, that it is the diagonalization of β, and hence $diag\big(gn(\beta)\big) = gn(\lambda)$. From this, by (1), follows

$$\mathbf{PA} \vdash \forall v_1(\mathbf{Diag}(\ulcorner\beta\urcorner, v_1) \leftrightarrow v_1 \doteq \ulcorner\lambda\urcorner) \tag{4}$$

That is, $\ulcorner\lambda\urcorner$ is the only closed term satisfying the open formula $\mathbf{Diag}(\ulcorner\beta\urcorner, v_1)$[8]. Simple logic then gives, from (3) and (4):

$$\mathbf{PA} \vdash \lambda \leftrightarrow \exists v_1(v_1 \doteq \ulcorner\lambda\urcorner \wedge \neg T(v_1)) \tag{5}$$

Since $\exists v_1(v_1 \doteq \ulcorner\lambda\urcorner \wedge \neg T(v_1))$ is equivalent to $\neg T(\ulcorner\lambda\urcorner)$, we have:

$$\mathbf{PA} \vdash \lambda \leftrightarrow \neg T\ulcorner\lambda\urcorner \qquad \square$$

This is the formal counterpart of the paradoxical Liar sentence: a sentence that is provably equivalent to a sentence saying that its code is not element of the extension of the truth predicate. "But note that $[\lambda]$ is produced by a simple diagonalization construction [...]; and the construction yields a theorem, not a paradox" (Smith,

[7] Whether this sentence "says of itself that it is not true" is not as obvious as one might think. For an insightful discussion about self-reference in arithmetic, see Halbach and Visser, (2014a,b).

[8] Note that (4) is equivalent to the conjunction of $\mathbf{PA} \vdash \mathbf{Diag}(\ulcorner\beta\urcorner, \ulcorner\lambda\urcorner)$ and $\mathbf{PA} \vdash \forall v_1(\neg(v_1 \doteq \ulcorner\lambda\urcorner) \rightarrow \neg\mathbf{Diag}(\ulcorner\beta\urcorner, v_1))$.

2013, p. 198). The "formal Liar paradox" arises if we want our theory of truth to prove the T-Schema $\varphi \leftrightarrow T\ulcorner\varphi\urcorner$ for all sentences $\varphi \in \mathscr{L}^t_{pa}$.

Yet, this is by no means necessary. Kripke (1975) proposes to give up the beloved T-Schema, constructing a partial model for the truth predicate, where both $\lambda \leftrightarrow T\ulcorner\lambda\urcorner$ and $\lambda \leftrightarrow \neg T\ulcorner\lambda\urcorner$ are neither true nor untrue, i.e. they are undefined. As indicated in the INTRODUCTION, I assume the reader being familiar with Kripke's *Outline*. I omit completely the presentation of his work. Here I shall just state two important features of MFP, described by Kripke (1975, p. 708) as "probably the most natural model for the intuitive concept of truth".

Fact 1.2. MFP verifies the metalinguistic T-Schema, i.e.: for all sentences $\varphi \in \mathscr{L}^t_{pa}$,

$$\langle\mathcal{M}, (E_\infty, A_\infty)\rangle \models_{\text{sk}} \varphi \;\Leftrightarrow\; \langle\mathcal{M}, (E_\infty, A_\infty)\rangle \models_{\text{sk}} T\ulcorner\varphi\urcorner$$

$$\langle\mathcal{M}, (E_\infty, A_\infty)\rangle \models_{\text{sk}} \neg\varphi \;\Leftrightarrow\; \langle\mathcal{M}, (E_\infty, A_\infty)\rangle \models_{\text{sk}} \neg T\ulcorner\varphi\urcorner$$

Fact 1.3. In MFP both the Liar sentence λ and the Truth-teller τ are undefined.

2 Towards a New Model

In this section I shall put forward the new response to the Liar antinomy. The gist of my proposal is that 'paradoxical' ought to be treated as a truth value. Liar sentences, according to the present suggestion, do not simply *lack* a truth value. They do possess one: they are *paradoxical*. As has been noted in the INTRODUCTION, the trigger of my considerations will be the difference between the Liar and the Truth-teller. The main goal is to construct a model within which (i) the difference between paradoxical and unparadoxical statements is detected, and (ii) every $\mathscr{L}^t_{pa}$-sentence φ has the same truth value as $T\ulcorner\varphi\urcorner$ (that's the metalinguistic T-Schema).

The plan is as follows: the next subsection contains philosophical arguments: I try to explain why Kripke's proposal is not sufficiently satisfactory as response to the Liar, and why, more generally, his MFP does not adequately model the truth predicate. In addition, I shall explain why 'paradoxical' should be treated as a truth value. The remaining subsections carry out this idea formally.

2.1 Why?

Without aiming to be censorious toward Kripke's proposal, but rather with the intention of further developing his elegant ideas, I think that his construction suffers from two inadequacies, which can (I hope) be removed. A first, minor problem his proposal is confronted with is that using the value 'undefined' for paradoxical sentences does not seem entirely adequate[9] – at least if we adhere to the original meaning attributed to it by Kleene, (1971). A second, major problem is that Kripke's MFP does not model the truth predicate in a satisfactory way. Let me elaborate these reasons in turn.

In both Kleene's logics (the *Strong* and the *Weak*)[10], the value 'undefined' ($\mathfrak{u}$) is not treated on a pair with 'true' (1) and 'false' (0): $\mathfrak{u}$ is not a third *truth value*[11]; it only represents formally the *lack* of truth values. Secondly, and more important for the present purposes, $\mathfrak{u}$ is open to "*arbitrariness* for a classical value": undefined sentences can turn out to be true or false, or can arbitrarily be declared true or false.

Less tersely: as is well known, Kleene introduced the new logics in the study of partial recursive functions, speaking of which he writes (Kleene, 1971, p. 334):

> if when $Q(x)$ is $\mathfrak{u}$, $Q(x) \vee R(x)$ receives the value 1, the decision must (in the general case) have been made in ignorance about $Q(x)$, and in the face of the possibility that, at some stage in the pursuit of the algorithm for $Q(x)$ later than the last one examined, $Q(x)$ might be found to be 1 or to be 0.

He goes on (*ibid.*, p. 335) to observe that 1, 0, and $\mathfrak{u}$ "must be susceptible of another meaning besides (i) 'true', 'false', 'undefined', namely (ii) 'true', 'false', 'unknown (or value immaterial)'. Here 'unknown' is a category, whose value we either do not know or choose for the moment to disregard; and it does not then exclude the other two possibilities 'true' or 'false' "[12].

My question now is: are paradoxical sentences like the Liar open to the same kind of arbitrariness for a classical value? Might these sentences turn out to be true, or false? Can we arbitrarily assign them a truth value? Hardly so. These sentences are paradoxical precisely because the assumption that they are true, or false, generates inconsistencies.

As already remarked, this is a minor problem. One might quite easily change the interpretation of $\mathfrak{u}$ and adjust it as pleased to paradoxes[13]. Nonetheless, the major problem continues to flutter: MFP does not model the truth predicate adequately, as it does not account for the difference between Liar and Truth-teller – this difference having its roots in a peculiarity of T. Let me make this claim precise, by first repeating that the difference between

[9] Some authors have suggested that paradoxes are overdefined (both true and false), and not underdefined (neither true nor false). See, for example, Dunn, (1969, 1976) and Priest, (1979).

[10] See Kleene, (1971, §64).

[11] Kripke (1975, fn 18) stresses the same point.

[12] Other philosophers have also suggested, as reported by van Fraassen, (1966, pp. 482-483), that sentences that are normally taken to be neither true nor false (for instance "the king of France is wise") "are 'don't cares' for ordinary purposes, and there is therefore no reason why we should not arbitrarily assign them some truth value".

[13] For example, Priest, (1979) introduced the so-called 'Logic of Paradox' (*LP*), which has the same truth tables as the Strong Kleene, but the interpretation of the third value is 'true and false', and it is, moreover, a designated value.

$ The sentence marked with a dollar is untrue.

and

€ The sentence marked with a euro is true.

is that one can more or less arbitrarily declare **€** true, or untrue, without stumbling on logical issues; on the contrary, the only way to declare **$** true, or untrue, requires the abandonment of an important principle about truth, i.e. that nothing is both true and untrue. Therefore, doing nothing more and nothing less than describing a simple state of affairs, we can state that

(Fact) *the truth predicate is such that, there are sentences that can* consistently *be in its extension or in its anti-extension; there are sentences that cannot.*

Every theory of truth ought to take **(Fact)** into account[14].

As a matter of fact, in a substantial portion of the *Outline*, Kripke shows how to categorise different kinds of sentence. A sentence is *paradoxical*, e.g., "if it has no truth value in *any* fixed point" (Kripke, 1975, p. 708)[15]. A sentence is *ungrounded* and *unparadoxical*, if it has a truth value in *some* fixed point, different from the minimal one – an example being the Truth-teller. He even emphasises that "the assignment of a truth value to [the Truth-teller] is *arbitrary*" (*ibid.*, p. 709)[16].

The reader might therefore ask, what the point of my objection is – Kripke *does* offer a way to distinguish between paradoxical and simply undefined sentences; Kripke *does* account for the difference between Liars and Truth-tellers. He surely does. But the point is that only within the *metatheory* one can implement that distinction. Only within an informal "metamodel" of the various fixed-point models are we able to differentiate between paradoxical and unparadoxical sentences. The minimal fixed point, which (*repetita iuvant*) is described by Kripke as "probably the most natural model for the intuitive concept of truth" (*ibid.*, p. 708), doesn't see the difference: in this model the Liar and the Truth-teller are both simply undefined.

If I am right, and if the difference between **$** and **€** is due to the peculiarity of T expressed by **(Fact)**[17], then I believe it is justified to maintain the Kripke's model

[14] A similar point is made by Gupta and Belnap, (1993, p. 100): "The essential thing about the Liar appears to be its instability under semantic evaluation: No matter what we hypothesize its value to be, semantic evaluation refutes our hypothesis. A theory of truth ought to capture *this* intuition. It should provide a way of distinguishing sentences that exhibit this behaviour from those that do not, and it should explain *why* certain sentences behave this way".

[15] Kripke considers only *consistent* fixed point, i.e. fixed point where $E \cap A = \emptyset$. So do I.

[16] Halbach, (2014, p. 196) observes that "Kripke's main contribution was not so much the construction of the smallest fixed point [...] but rather his classification of the different consistent fixed points and the discussion of their use for discriminating between ungrounded sentences, paradoxical sentences, and so on".

[17] Are there any other predicates which are akin to T in this respect? One is there for sure: the predicate "is heterological" introduced by Kurt Grelling and Leonard Nelson (see Grelling and Nelson, 1907). In a parallel work, I am trying to extend the solution presented here to handle the Grelling-Nelson paradox too.

is not quite accurate. I believe it is justified to maintain that we should try to find a way to improve it.

Some suggestions have already been made: it is what McGee, (1991, pp. 110-111) calls a 'liberalisation of Kripke's construction', which allows extension and anti-extension of T to overlap. This requires a replacement of a 3-valued logic with a 4-valued logic having both truth value gaps and truth value gluts. The logical-mathematical properties of such a liberalisation have been studied by Woodruff, (1984)[18]. Such systems are of great interest for dialetheists[19]. But for those who do not believe that something can be ever both true and false, they are of little help. I am one of those, and additionally I really do not believe that declaring the Liar both true and false can represent any kind of solution to the paradox. It seems to me that the paradox *is* precisely that some sentence should be both true and false. I can't digress, however, to discuss dialetheism – intriguing though it might be.

2.2 How?

Although I am not an advocate of dialetheism, I subscribe Visser's words, when he says that "[o]ne attractive feature of four valued logic for the study of the Liar Paradox is the possibility of making certain intuitive distinctions [that is: the distinction between Liars and Truth-tellers. L.C.] *within one single model*" (Visser, 1984, pp. 181-182). And that is why I am about to introduce a new 4-valued logic, whose values are: true, false, paradoxical, and undefined. "Why 'paradoxical'?" – the reader might ask. To properly answer this question, I first need to introduce the idea underlying the new interpretation of T.

We all agree (I venture) that an adequate interpretation of the truth predicate ought to have an extension E and an anti-extension A. Now, since (i) I do not want Liar sentences to simply lay outside $E \cup A$ with Truth-teller sentences, and since (ii) I do not want E and A to overlap, I propose to extend Kripke's interpretation of T by adding a third set to it, which will contain those (codes of) sentences that, as stated in **(Fact)**, cannot consistently be contained in E or in A. I shall call this third set (due to lack of imagination) X. In particular: (E, A, X) will be the interpretation of T, the interpretation of $\mathscr{L}_{pa}$ remaining as before, i.e. we let $\mathcal{M}$ be the standard interpretation of $\mathscr{L}_{pa}$. Consequently, $\langle \mathcal{M}, (E, A, X) \rangle$ will be the interpretation of $\mathscr{L}^t_{pa}$ with, informally:

(i) $E = \{gn(\varphi) \mid \varphi \text{ is true}\}$; $A = \{gn(\varphi) \mid \varphi \text{ is untrue}\}$; $X = \{gn(\varphi) \mid \varphi \text{ is paradoxical}\}$;

[18] See also Visser, (1984).

[19] *Dialetheism*, roughly, is the view that there are true contradictions, and a full exposition of it would involve a great deal of technical material that we will not go into here. See Priest and Berto, (2013) for an overview.

(ii) $E \cap A = \emptyset$, $E \cap X = \emptyset$, $A \cap X = \emptyset$;

(iii) $E \cup A \cup X \neq \mathbb{N}$.

And so now the question arises, what truth value sentences like $T\ulcorner\varphi\urcorner$ should have, whenever $gn(\varphi) \in X$. The answer suggested here, unsurprisingly, is that they are paradoxical. Hence, the reason why I am proposing to take 'paradoxical' as a truth value is that I think the best way to formalise **(Fact)** is having a threefold interpretation of T, with extension, anti-extension, and paradox-set. Accordingly, exactly as though we were allowing E and A to overlap, a fourth truth value is needed. And no value but 'paradoxical' seems to properly suit the paradox-set X.

Now, to carry out this project formally, there are above all three things to be done: first, we need a new 4-valued logic to handle the value 'paradoxical'; second, we need rules determining whether a sentence is true, false, paradoxical, or undefined in the partial model $\langle \mathcal{M}, (E, A, X)\rangle$; third, we need a formal definition of (E, A, X).

2.3 The New Logic

2.3.1 Truth Values and their Structure

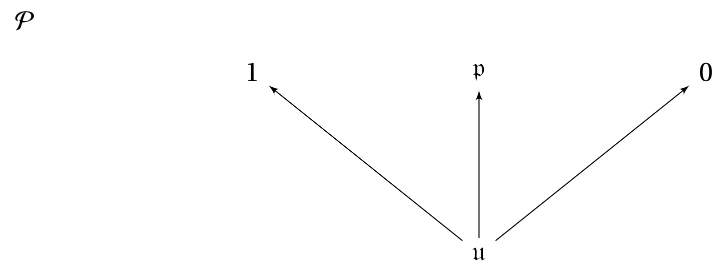
Let $\mathbb{C}$ be the class of connectives of classical propositional logic. The new 4-valued logic is defined by the structure:

$$\begin{aligned} \mathcal{W} &= \{1, 0, \mathfrak{p}, \mathfrak{u}\} \\ \mathcal{D} &= \{1\} \\ C &= \{f_c \mid c \in \mathbb{C}\} \end{aligned}$$

where $\mathcal{W}$ is the set of truth values (true, false, paradoxical, undefined), $\mathcal{D}$ the set of the sole designated value, C the set of truth functions: for every connective $c \in \mathbb{C}$, f_c is the corresponding truth function. That is: if $c \in \mathbb{C}$ is an n-place connective, f_c is a n-place function with inputs and outputs in $\mathcal{W}$.

As usual, one might order the element of $\mathcal{W}$ by the relation $\leq$. Since $\mathfrak{u}$ represents the lack of truth values, we will have: $\mathfrak{u} \leq 1$; $\mathfrak{u} \leq 0$; $\mathfrak{u} \leq \mathfrak{p}$. The decision to be made concerns the new value $\mathfrak{p}$. There are three possibilities. One might argue that 'paradoxical' represents some sense of 'overdefined', in which case we would have $1 \leq \mathfrak{p}$, $0 \leq \mathfrak{p}$. Or one might say that, like $\mathfrak{u}$, $\mathfrak{p}$ stands for another case of 'underdefined', in which case we would have $\mathfrak{p} \leq 1$ and $\mathfrak{p} \leq 0$. Alternatively, one might say, as I shall do here, that it is neither 'overdefined', nor 'underdefined', whence we have: 1, 0, and $\mathfrak{p}$ are not comparable.

This yields a structure $\mathcal{P} = \langle \mathcal{W}, \leq \rangle$, which can be pictured thus:



$\mathcal{P}$ is a poset (partially ordered set), since the ordering $\leq$ on $\mathcal{W}$ is a reflexive, transitive, and antisymmetric binary relation.

Definition 2.1 (Consistency and ccpo)**.** Let $P = \langle D, \leq \rangle$ be a poset. Following Visser, (1984, pp. 184-185), define

(a) A subset $A \subseteq D$ is *consistent* iff each $\{x, y\} \subseteq A$ has an upper bound in D.

(b) P is a *complete, coherent partial order* (ccpo), iff every consistent subset $A \subseteq D$ has a supremum.

Proposition 2.2. $\mathcal{P}$ is a ccpo.

Proof. It is easily verified that each consistent pair of elements $\{\mathfrak{u}, 0\}$, $\{\mathfrak{u}, 1\}$, $\{\mathfrak{u}, \mathfrak{p}\} \subseteq \mathcal{W}$ has a supremum in $\mathcal{W}$ (respectively: 0, 1, $\mathfrak{p}$)[20]. □

2.3.2 Truth Tables and Valuation Function

Instead of defining truth functions singularly[21], I shall for simplicity use the truth tables and I shall write the simple connectives $\neg$, $\vee$, $\wedge \ldots$ instead of $f_\neg$, $f_\vee$, $f_\wedge \ldots$ I also write explicitly conjunction, conditional, and biconditional, although they are defined as usual through negation and disjunction.

$\neg$	
1	0
0	1
$\mathfrak{p}$	$\mathfrak{p}$
$\mathfrak{u}$	$\mathfrak{u}$

$\vee$	1	0	$\mathfrak{p}$	$\mathfrak{u}$
1	1	1	1	1
0	1	0	$\mathfrak{p}$	$\mathfrak{u}$
$\mathfrak{p}$	1	$\mathfrak{p}$	$\mathfrak{p}$	$\mathfrak{u}$
$\mathfrak{u}$	1	$\mathfrak{u}$	$\mathfrak{u}$	$\mathfrak{u}$

$\wedge$	1	0	$\mathfrak{p}$	$\mathfrak{u}$
1	1	0	$\mathfrak{p}$	$\mathfrak{u}$
0	0	0	0	0
$\mathfrak{p}$	$\mathfrak{p}$	0	$\mathfrak{p}$	$\mathfrak{u}$
$\mathfrak{u}$	$\mathfrak{u}$	0	$\mathfrak{u}$	$\mathfrak{u}$

[20] Gupta and Belnap, (1993, §2C) study the mathematical properties of complete coherent partial orders, which turn out to be useful in investigating truth in three-valued languages.

[21] For instance:

$$f_\neg(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{if } x = 1 \\ \mathfrak{p} & \text{if } x = \mathfrak{p} \\ \mathfrak{u} & \text{if } x = \mathfrak{u} \end{cases}$$

$\rightarrow$	1	0	$\mathfrak{p}$	$\mathfrak{u}$
1	1	0	$\mathfrak{p}$	$\mathfrak{u}$
0	1	1	1	1
$\mathfrak{p}$	1	$\mathfrak{p}$	$\mathfrak{p}$	$\mathfrak{u}$
$\mathfrak{u}$	1	$\mathfrak{u}$	$\mathfrak{u}$	$\mathfrak{u}$

$\leftrightarrow$	1	0	$\mathfrak{p}$	$\mathfrak{u}$
1	1	0	$\mathfrak{p}$	$\mathfrak{u}$
0	0	1	$\mathfrak{p}$	$\mathfrak{u}$
$\mathfrak{p}$	$\mathfrak{p}$	$\mathfrak{p}$	$\mathfrak{p}$	$\mathfrak{u}$
$\mathfrak{u}$	$\mathfrak{u}$	$\mathfrak{u}$	$\mathfrak{u}$	$\mathfrak{u}$

Do the tables suit our intuitions about paradoxality? I will discuss this question below, in Discussion 2.4. But before that, let me define the valuation function $\mathcal{V}_{\langle \mathcal{M},(E,A,X)\rangle} : \mathscr{L}^t_{pa} \longrightarrow \{1, 0, \mathfrak{p}, \mathfrak{u}\}$. For the sake of readability, I shall write $\mathcal{V}$ instead of $\mathcal{V}_{\langle \mathcal{M},(E,A,X)\rangle}$.

(a) For atomic $\mathscr{L}_{pa}$-sentences:

$$\mathcal{V}(\varphi) = \begin{cases} 1 & \text{if} \quad \mathcal{M} \models \varphi \\ 0 & \text{if} \quad \mathcal{M} \models \neg\varphi \end{cases}$$

(b) For atomic $\mathscr{L}^t_{pa}$-sentences $T(\mathbf{n})$:

$$\mathcal{V}(T(\mathbf{n})) = \begin{cases} 1 & \text{if} \quad n \in E \\ 0 & \text{if} \quad n \in A \\ \mathfrak{p} & \text{if} \quad n \in X \\ \mathfrak{u} & \text{if} \quad n \notin E \cup A \cup X \end{cases}$$

(c)

$$\mathcal{V}(\neg\varphi) = \begin{cases} 1 & \text{if} \quad \mathcal{V}(\varphi) = 0 \\ 0 & \text{if} \quad \mathcal{V}(\varphi) = 1 \\ \mathfrak{p} & \text{if} \quad \mathcal{V}(\varphi) = \mathfrak{p} \\ \mathfrak{u} & \text{if} \quad \mathcal{V}(\varphi) = \mathfrak{u} \end{cases}$$

(d)

$$\mathcal{V}(\exists v_i \varphi(v_i)) = \begin{cases} 1 & \text{if} \quad \exists n \in \mathbb{N}\,\big(\mathcal{V}\big(\varphi(\mathbf{n})\big) = 1\big) \\ 0 & \text{if} \quad \forall n \in \mathbb{N}\,\big(\mathcal{V}\big(\varphi(\mathbf{n})\big) = 0\big) \\ \mathfrak{p} & \text{if} \quad \text{(see below)} \\ \mathfrak{u} & \text{if} \quad \text{(see below)} \end{cases}$$

The definition for compound sentences containing connectives is given on the basis of the valuation scheme. The definition for quantified sentences is more intricate, so let me explain the process that brought me at the definition presented below.

When does a sentence beginning with a quantifier have semantic value $\mathfrak{p}$? The answer to this question is crucial, since the various paradoxical sentences are exactly quantified sentences. More precisely, they have the form $\exists v_0(\varphi(v_0) \wedge \neg T(v_0))$, where the code of the sentence is the *only* object satisfying the formula $\varphi(v_0)$, so that for all other numbers n, $\varphi(\mathbf{n})$ is false.

Now, the semantic rules determining when a quantified sentence is true or false can be borrowed from the Strong Kleene semantics adopted by Kripke – as I already did in (d). The problem is that a companion definition for paradoxality, namely

$$\mathcal{V}(\exists v_i \varphi(v_i)) = \mathfrak{p} \text{ iff } \exists n \in \mathbb{N} \left(\mathcal{V}\big(\varphi(\mathbf{n})\big) = \mathfrak{p}\right)$$

is evidently inadequate, since for $\varphi(v_0) \equiv T(v_0)$ there is indeed a $n \in \mathbb{N}$ such that $\mathcal{V}(T(\mathbf{n})) = \mathfrak{p}$, but the sentence "something is true" is not paradoxical. Certainly, nonetheless, the condition that there must be a $n \in \mathbb{N}$, such that $\mathcal{V}(\varphi(\mathbf{n})) = \mathfrak{p}$, is a necessary condition – though not sufficient.

A second thought might be

$$\mathcal{V}(\exists v_i \varphi(v_i)) = \mathfrak{p} \text{ iff } \forall n \in \mathbb{N} \left(\mathcal{V}\big(\varphi(\mathbf{n})\big) = \mathfrak{p}\right)$$

This also does not work, since it would not make λ, as presented in Lemma 1.1, paradoxical. Recall that λ is the sentence $\exists v_0(v_0 \doteq \ulcorner\beta\urcorner \wedge \beta(v_0))$. But the formula $v_0 \doteq \ulcorner\beta\urcorner \wedge \beta(v_0)$ is not always paradoxical. Quite the opposite, for each $n \neq gn(\beta)$, it is false. Certainly, nonetheless, the condition that $\mathcal{V}(\varphi(\mathbf{n})) = \mathfrak{p}$ for all $n \in \mathbb{N}$ is a sufficient condition – though not necessary.

Combining now sufficient and necessary conditions, I shall propose the following definition:

(d)

$$\mathcal{V}(\exists v_i \varphi(v_i)) = \begin{cases} 1 & \text{if} \quad \exists n \in \mathbb{N} \left(\mathcal{V}\big(\varphi(\mathbf{n})\big) = 1\right) \\ 0 & \text{if} \quad \forall n \in \mathbb{N} \left(\mathcal{V}\big(\varphi(\mathbf{n})\big) = 0\right) \\ \mathfrak{p} & \text{if} \quad \exists n \in \mathbb{N} \left(\mathcal{V}\big(\varphi(\mathbf{n})\big) = \mathfrak{p}\right) \wedge \\ & \qquad \forall m \in \mathbb{N} \left(\mathcal{V}\big(\varphi(\mathbf{m})\big) = \mathfrak{p} \vee \mathcal{V}\big(\varphi(\mathbf{m})\big) = 0\right) \\ \mathfrak{u} & \text{if} \quad \exists n \in \mathbb{N} \left(\mathcal{V}\big(\varphi(\mathbf{n})\big) = \mathfrak{u}\right) \wedge \\ & \qquad \neg\exists n \in \mathbb{N} \left(\mathcal{V}\big(\varphi(\mathbf{n})\big) = 1\right) \end{cases}$$

The universal quantifier is defined as usual thus:

$$\forall v_i(\varphi(v_i)) :\leftrightarrow \neg\exists v_i \neg(\varphi(v_i))$$

To complete the rough description of the new logic, let me add that validity is defined in terms of truth preservation: an inference from Σ to φ is valid iff if for each $\psi \in \Sigma$, $\mathcal{V}(\psi) = 1$ (the sole designated value), then $\mathcal{V}(\varphi) = 1$.

It would be interesting to compare this logic with some of the 4-valued logics already studied in the literature. But this demands a larger discussion than is possible here. I shall only make one quick remark:

Remark 2.3. Disjunctive syllogism (from $\varphi \vee \psi$ and $\neg\varphi$ infer ψ) is not valid in the 4-valued logic called *first degree entailment*. This is due to the fact that the designated values of this logic are 1 and $\mathfrak{b}$ (= 'both'). As an example, assume that $\varphi = \mathfrak{b}$ and $\psi = 0$; then $\neg\varphi, \varphi \vee \psi \not\models_{FDE} \psi$, since both $\neg\varphi$ and $\varphi \vee \psi$ are designated (namely $\mathfrak{b}$), but ψ undesignated. On the contrary, it is easily verified that in the logic just sketched disjunctive syllogism is valid, for the only designated value is 1[22]. ■

Let us now turn on the truth tables. They are (i) *truth-functional*, in the sense that the value of a compound is a function of the values of its immediate components; (ii) *normal*, in the sense that the value of a compound is determined by the classical rules whenever the components have classical value; (iii) *monotonic*, for they preserve the relevant order.

Behind them there are four simple thoughts: first, they are an extension of the Strong Kleene (K_3) – in fact, whenever no component is $\mathfrak{p}$, they are exactly as K_3; second, the value 'paradoxical' behaves exactly like $\mathfrak{u}$ in connection with 1 or 0; third, the connection of $\mathfrak{u}$ and $\mathfrak{p}$ is always undefined; fourth, like K_3, they let classical logic be our guiding light, whenever we have "enough classical information". Classical logic, for instance, tells us that a conjunction is false whenever at least one conjunct is false. Accordingly, if a conjunction has a false conjunct, the whole sentence becomes false, *independently* from the value of the other conjunct.

Discussion 2.4. Do the tables suit our intuitions about paradoxality? Besides the case of negation, it is hard to determine, since we do not utter, in the everyday life, many compound sentences containing paradoxes as components. I shall thus make no claim to the optimality of the chosen scheme. By way of an example, however, consider:

♣ The part before the comma of the sentence marked with a clubs sign
is untrue, or 0 = 0 [formalisable as $\lambda \vee \mathbf{0} \doteq \mathbf{0}$].

♠ The part before the comma of the sentence marked with a spade sign
is untrue, or 0 = 1 [formalisable as $\lambda \vee \mathbf{0} \doteq \mathbf{1}$].

[22] Whether the disjunctive syllogism is a plus or a minus is controversial. See Priest, (2006, p. 154) for a brief discussion.

Although these sentences are highly artificial, they ought to be taken into account when working with formal languages. The former seems to be true, simply because it is a disjunction containing a true disjunct. And, as already remarked, classical logic ought to be our guiding light, whenever classical information is enough.

The second sentence might give some troubles. According to the tables, it is paradoxical and this choice is prompted by two considerations. The first: the sentence is surely neither true nor undefined. Now, if we assume that ♠ is untrue, then both disjuncts have to be untrue (this implication presupposes, again, to follow classical logic as far as possible). But the part before the comma is untrue if, and only if, it is true. The second consideration: it creates a parallel with K_3 and with the work of Kripke. In fact, within Kripke's framework, ♠ would be undefined, and undefined is the value ascribed to λ. Since in the new framework λ has a new truth value, the whole sentence does get a new value as well. Nonetheless, the idea that the sentence is assigned the value of λ is preserved. ■

We can now move on to the last part of this section.

2.4 The New Interpretation of T

To begin with, I shall exploit Kripke's construction of Mfp: in the new interpretation of T, E and A will be identical to E_∞ and A_∞ (the extension and the anti-extension of T in Mfp). Of interest is the definition of the paradox-set and the differentiation between paradoxical and ungrounded-and-unparadoxical sentences. Before I begin, a quick remark on the choice of letting E and A be identical to E_∞ and A_∞. Whereas Kripke maintains that the minimal fixed point is *probably* the most natural model for the intuitive concept of truth, I go a bit further: Mfp *is* the most natural model for the ordinary truth predicate[23]. In a longer philosophical work I would have defended this claim. But limits in space urges us to move on to the formal definition of X.

Recall the way Kripke defines paradoxical sentences, namely: a sentence is paradoxical if, and only if, it does not have a truth value in any (consistent) fixed point, whereas a sentence is ungrounded and unparadoxical iff it has a truth value in some fixed point, different from the minimal one. Now, one might be tempted to formalise Kripke's characterisation word for word, defining X as the set of all (codes of) sentences that are undefined in every fixed point. Such a definition would make all Liar sentences paradoxical, and all Truth-teller sentences unparadoxical – and these are indeed two desiderata of the new model. But an unpleasant consequence would

[23] My statement ranges over kripkean models.

derive from it. Let τ be a Truth-teller[24]. If I defined X as above, the following, e.g., would hold in the new model:

$$\mathcal{V}\big((\tau \wedge \neg\tau) \vee \lambda\big) = \mathfrak{u} \qquad \mathcal{V}\big(T\ulcorner(\tau \wedge \neg\tau) \vee \lambda\urcorner\big) = \mathfrak{p} \tag{6}$$

$$\mathcal{V}\big((\tau \vee \neg\tau) \wedge \lambda\big) = \mathfrak{u} \qquad \mathcal{V}\big(T\ulcorner(\tau \vee \neg\tau) \wedge \lambda\urcorner\big) = \mathfrak{p} \tag{7}$$

As (6)-left never gets a truth value in any fixed point, it should be element of X, so that (6)-right would be paradoxical in $\langle\mathcal{M}, (E, A, X)\rangle$. Yet, (6)-left is undefined in $\langle\mathcal{M}, (E, A, X)\rangle$, because $\tau \wedge \neg\tau$ is undefined and λ paradoxical[25]. Similarly for (7).

Therefore, I cannot define X this way, for that would mean abandoning the prospect of constructing a model where every sentence φ has the same truth value as $T\ulcorner\varphi\urcorner$. I shall hence posit a different definition.

Kripke, (1975, p. 701) makes the following example: "Suppose we are explaining the word 'true' to someone who does not yet understand it. We may say that we are entitled to assert (or deny) of any sentence that it is true precisely under the circumstances when we can assert (or deny) the sentence itself". Following this example, I would suggest:

> we are entitled to assert of any sentence that it is paradoxical under the circumstances when we cannot assert the sentence itself, without being led to assert that it is untrue.

This informal picture is obviously meant to characterise truth-related paradoxes, like the Liar or like the example from Kripke, (1975, p. 691), which involves a kind of cross-reference between statements: Jones says

(I) Most of Nixon's assertions about Watergate are false.

Suppose now that Nixon's assertions about Watergate are evenly balanced between the true and the false, except for one problematic case:

(II) Everything Jones says about Watergate is true.

Suppose, in addition, that **(I)** is the only statement of Jones about Watergate. It is easy to verify that we cannot assert **(I)** (or **(II)**), without being led to assert that it is untrue: If we assert **(I)**, we are implying that **(II)** is untrue. But this implies that **(I)** is untrue. Similarly if we deny **(I)**[26].

[24] Whereas Liar sentences have the form $\exists v_0(\varphi(v_0) \wedge \neg T(v_0))$, Truth-teller sentences are $\exists v_0(\varphi(v_0) \wedge T(v_0))$. In both cases, the code of the sentence in the only number satisfying the formula $\varphi(v_0)$.

[25] Notice that, although I haven't yet shown it in details, τ is undefined and λ paradoxical according to the definition of quantified sentences above. See infra, Proposition 3.1, for details.

[26] Paying attention at some details, also Yablo's paradoxical sequence (see Yablo, 1993) could be described in the same manner.

Let me now turn the informal description of paradoxical sentences into a formal definition. I shall define the set X inductively. After that, I shall explain how the formal definition relates to the informal characterisation.

First, let $\zeta(n, S)$ abbreviate

(i) $n = gn(\varphi) ⩕ \mathbf{PA} \vdash \varphi \leftrightarrow \neg T\ulcorner\varphi\urcorner$; or

(ii) $n = gn(\neg\varphi) ⩕ gn(\varphi) \in S$; or

(iii) $n = gn(\varphi \vee \psi) ⩕ \quad (gn(\varphi) \in S ⩖ gn(\psi) \in S) ⩕$
$\big((gn(\varphi) \in S \Rightarrow gn(\psi) \in S \cup A) ⩕ (gn(\psi) \in S \Rightarrow gn(\varphi) \in S \cup A)\big)$; or

(iv) $n = gn(\varphi \wedge \psi) ⩕ \quad (gn(\varphi) \in S ⩖ gn(\psi) \in S) ⩕$
$\big((gn(\varphi) \in S \Rightarrow gn(\psi) \in S \cup E) ⩕ (gn(\psi) \in S \Rightarrow gn(\varphi) \in S \cup E)\big)$[27]; or

(v) $n = gn(\exists v_i \varphi(v_i)) ⩕$
$\exists m \in \mathbb{N}\,\big(gn(\varphi(\mathbf{m})) \in S\big) ⩕ \forall k \in \mathbb{N}\,\big(gn(\varphi(\mathbf{k})) \in S \cup A\big)$; or

(vi) $n = gn(T(\mathbf{m})) ⩕ m \in S$.

This gives rise to an operator Γ on the powerset of natural numbers, which is monotone. It is well known that monotone operators on $\mathcal{P}(\mathbb{N})$ have fixed points. The minimal one will be our set X.

Definition 2.5 (Paradox Operator)**.** The paradox operator $\Gamma : \mathcal{P}(\mathbb{N}) \longrightarrow \mathcal{P}(\mathbb{N})$ is a function on the powerset of $\mathbb{N}$, defined thus:

$$\Gamma(S) = \{n \mid \zeta(n, S)\}$$

Example 2.6. Let $S_0 = \{gn(\mathbf{0} \doteq \mathbf{0})\}$. Then $\Gamma(S_0)$ will first of all contain all n, such that $n = gn(\varphi)$ and $\mathbf{PA} \vdash \varphi \leftrightarrow \neg T\ulcorner\varphi\urcorner$. Moreover, by condition (ii), it will contain all $n = gn(\neg\varphi)$ such that $gn(\varphi) \in S_0$. Now, since the only $gn(\varphi) \in S_0$ is $gn(\mathbf{0} \doteq \mathbf{0})$, $gn(\neg(\mathbf{0} \doteq \mathbf{0}))$ will be the only (code of) sentence obtained through condition (ii); by condition (iii), $\Gamma(S_0)$ will contain all $n = gn(\varphi \vee \psi)$ such that $gn(\varphi) \in S_0$ or $gn(\psi) \in S_0 \ldots$ and so forth. In our case, since $S_0 = \{gn(\mathbf{0} \doteq \mathbf{0})\}$, $\Gamma(S_0)$ will contain sentences like $gn(\mathbf{0} \doteq \mathbf{0} \vee \mathbf{0} \doteq \mathbf{0})$ (because $gn(\mathbf{0} \doteq \mathbf{0}) \in S_0$), or $gn(\mathbf{0} \doteq \mathbf{0} \vee \mathbf{1} \doteq \mathbf{2})$ (because $gn(\mathbf{1} \doteq \mathbf{2}) \in A$) and so on. Obviously, it will not contain sentences like $gn(\mathbf{0} \doteq \mathbf{0} \vee \mathbf{1} \doteq \mathbf{1})$ (because $gn(\mathbf{1} \doteq \mathbf{1}) \notin S_0 \cup A$).

Notice that, since $gn(\lambda) \notin S_0 \cup A \cup E$, sentences like $\lambda \vee \psi$, $\lambda \wedge \psi$ will not be in $\Gamma(S_0)$, regardless of the ψ. However, being λ provably equivalent with $\neg T\ulcorner\lambda\urcorner$, it will

[27] The reader knows that the conjunction symbol is not part of the official language. I include it anyway, to obtain a clearer overview.

be, according to condition (i), in $\Gamma(S_0)$. Consequently, $gn(\lambda \vee \psi)$ or $gn(\lambda \wedge \psi)$ will be in $\Gamma(\Gamma(S_0))$, whenever ψ respects the conditions imposed by the definition. ■

In section 2.2, I have claimed that $X \cap E = \emptyset$ and that $X \cap A = \emptyset$. Clearly, if we start the iteration of Γ as shown in Example 2.6, we will not obtain this result. On the other hand, I have also claimed that X is the least fixed point of Γ, which is obtained by starting the sequence with $S_0 = \emptyset$. To show that Γ has a least fixed point, it suffices to show that it is a monotone function on $\mathcal{P}(\mathbb{N})$. After having shown the monotonicity, it will follow from general theory of inductive definitions that Γ has a least fixed point.

Lemma 2.7 (Monotonicity)**.** Γ is monotone. That is: for all $S_i, S_j \in \mathcal{P}(\mathbb{N})$,

$$S_i \subseteq S_j \Rightarrow \Gamma(S_i) \subseteq \Gamma(S_j)$$

Proof. Let $S_1 \subseteq S_2$ and assume, towards a contradiction

$$\exists n \in \mathbb{N}(n \in \Gamma(S_1) \wedge\!\!\wedge\, n \notin \Gamma(S_2)) \tag{8}$$

Let k be a number obtained through existential elimination. From the assumption that $k \in \Gamma(S_1)$ follows:

(i) $k = gn(\varphi) \wedge\!\!\wedge\, \mathbf{PA} \vdash \varphi \leftrightarrow \neg T\ulcorner\varphi\urcorner$; or

(ii) $k = gn(\neg\varphi) \wedge\!\!\wedge\, gn(\varphi) \in S_1$; or

(iii) $k = gn(\varphi \vee \psi) \wedge\!\!\wedge\, (gn(\varphi) \in S_1 \vee\!\!\vee\, gn(\psi) \in S_1) \wedge\!\!\wedge$
$\big((gn(\varphi) \in S_1 \Rightarrow gn(\psi) \in S_1 \cup A) \wedge\!\!\wedge\, (gn(\psi) \in S_1 \Rightarrow gn(\varphi) \in S_1 \cup A)\big)$; or

(iv) $k = gn(\varphi \wedge \psi) \wedge\!\!\wedge\, (gn(\varphi) \in S_1 \vee\!\!\vee\, gn(\psi) \in S_1) \wedge\!\!\wedge$
$\big((gn(\varphi) \in S_1 \Rightarrow gn(\psi) \in S_1 \cup E) \wedge\!\!\wedge\, (gn(\psi) \in S_1 \Rightarrow gn(\varphi) \in S_1 \cup E)\big)$; or

(v) $k = gn(\exists v_i \varphi(v_i)) \wedge\!\!\wedge$
$\exists n \in \mathbb{N}\big(gn(\varphi(\mathbf{n})) \in S_1\big) \wedge\!\!\wedge\, \forall m \in \mathbb{N}\big(gn(\varphi(\mathbf{m})) \in S_1 \cup A\big)$; or

(vi) $k = gn(T(\mathbf{n})) \wedge\!\!\wedge\, n \in S_1$.

It can be shown that each of (i)-(vi) implies that $k \in \Gamma(S_2)$.

If (i), then trivially $k \in \Gamma(S_2)$.

If (ii), as $S_1 \subseteq S_2$, $gn(\varphi) \in S_2$, and hence $gn(\neg\varphi) \in \Gamma(S_2)$.

For (iii), let me proceed slowly, step by step. First of all, I have to show that:

$$(iii) \quad \Rightarrow \quad (iii)[S_2/S_1] \tag{9}$$

That is: if (iii) is true (viz. if $k = gn(\varphi \vee \psi) \mathbin{⩕}(gn(\varphi) \in S_1 \mathbin{⩔} gn(\psi) \in S_1) \mathbin{⩕} \ldots$), then also $(iii)[S_2/S_1]$ is verified. Now, since we are assuming (iii), we are assuming in particular that $(gn(\varphi) \in S_1 \mathbin{⩔} gn(\psi) \in S_1)$, which implies the second conjunct of $(iii)[S_2/S_1]$, i.e. $(gn(\varphi) \in S_2 \mathbin{⩔} gn(\psi) \in S_2)$ (the first conjunct holds anyway). In order to show the third conjunct, I shall conduct a proof by cases: exploiting the assumption that $(gn(\varphi) \in S_1 \mathbin{⩔} gn(\psi) \in S_1)$, I shall show that both implies the third conjunct of $(iii)[S_2/S_1]$. In symbols:

$$\begin{gathered}(gn(\varphi) \in S_1 \mathbin{⩔} gn(\psi) \in S_1) \Rightarrow \\ \big((gn(\varphi) \in S_2 \Rightarrow gn(\psi) \in S_2 \cup A) \mathbin{⩕} (gn(\psi) \in S_2 \Rightarrow gn(\varphi) \in S_2 \cup A)\big)\end{gathered} \tag{10}$$

Assume first that $gn(\varphi) \in S_1$. Then $gn(\varphi) \in S_2$, and therefore the second conjunct of (10) is true. To show the first conjunct, notice that from the assumption that $gn(\varphi) \in S_1$ follows that $gn(\psi) \in S_1 \cup A$ and hence that $gn(\psi) \in S_2 \cup A$. This verifies the first conjunct of (10) and concludes the first part of the proof by cases, that is to say: if $gn(\varphi) \in S_1$, then the third conjunct of $(iii)[S_2/S_1]$ is true.

The second part of the proof by cases, which involves the assumption that $gn(\psi) \in S_1$, is exactly the same (*mutatis mutandis*, of course). Hence, if (iii), then $(iii)[S_2/S_1]$ and therefore $k \in \Gamma(S_2)$.

If (iv), then it suffices to substitute E for A in the argument above.

If (v), then there is a $n \in \mathbb{N}$, such that $gn(\varphi(\mathbf{n})) \in S_2$ and for all $m \in \mathbb{N}$, $gn(\varphi(\mathbf{m})) \in S_2 \cup A$. Therefore $gn(\exists v_0 \varphi(v_0)) \in \Gamma(S_2)$.

If (vi), then $n \in S_2$ and hence $gn(T(\mathbf{n})) \in \Gamma(S_2)$.

(8) is therefore false, and the monotonicity of Γ is proved. □

Since Γ is a monotone operator on $\mathcal{P}(\mathbb{N})$, it has a least fixed point.

Lemma 2.8 (Fixed Point)**.** Γ has a minimal fixed point, i.e. there is a set S such that $\Gamma(S) = S$, and for all $S' = \Gamma(S')$, $S \subseteq S'$.

Proof Sketch. Every monotone function $\pi : P \longrightarrow P$ on an inductive poset P[28] has a (unique) least fixed point. Since the paradox operator Γ is a monotone function on the power set of natural numbers, and since $\mathcal{P}(\mathbb{N})$ is an inductive poset, Γ has a least fixed point[29]. □

[28] A poset P is *inductive* (or *chain-complete*) if every chain $S \subseteq P$ has a least upper bound. (Moschovakis, 2006, Def. 6.10, p. 75).

[29] See Moschovakis, (2006, §§6-7), and Moschovakis, (1974, pp. 6-8) for details. The former contains an extensive, yet accessible, analysis of fixed points in general. The latter is a study of inductive definitions.

The informal description of paradoxical sentences stated above is captured by the first clause of the formal definition, viz. $n = gn(\varphi) \mathbin{\wedge\!\!\wedge} \mathbf{PA} \vdash \varphi \leftrightarrow \neg T\ulcorner\varphi\urcorner$. It makes sure that the "atomic" paradoxical sentences are elements of X. These sentences, evidently, are not atomic in the usual sense. Nevertheless, they are atomic in the sense that they are the minimum required to yield a paradox. All other clauses are meant to avoid the problem which would have followed from a definition in "Kripke-style"[30]. In other words: their goal is, on the basis of the truth tables presented in § 2.3.2, to assure that in the new model a sentence φ is paradoxical if, and only if, $T\ulcorner\varphi\urcorner$ is paradoxical too. A proof of this claim is contained in the following and last section, which contain the main theorem of the paper.

3 Analysis of the New Model

Let us check whether the model constructed thus far adequately models the truth predicate, and whether it improves the kripkean Mfp. First of all, I will show that the Liar gets assigned value $\mathfrak{p}$. Thereafter, I shall prove that the new model verifies the metalinguistic T-Schema.

Proposition 3.1. In $\langle \mathcal{M}, (E, A, X)\rangle$ both λ and $\neg T\ulcorner\lambda\urcorner$ are paradoxical.

Proof. I follow the notation of Lemma 1.1. Since λ is provably equivalent (in **PA**) with $\neg T\ulcorner\lambda\urcorner$, it follows that $gn(\lambda) \in X$ and therefore $\mathcal{V}(T\ulcorner\lambda\urcorner) = \mathfrak{p}$ iff $\mathcal{V}(\neg T\ulcorner\lambda\urcorner) = \mathfrak{p}$.

To prove that $\mathcal{V}(\lambda) = \mathfrak{p}$, as λ is a sentence beginning with a quantifier, namely

$$\exists v_0\big(\underbrace{v_0 \doteq \ulcorner\beta\urcorner \wedge \exists v_1(\mathbf{Diag}(v_0, v_1) \wedge \neg T(v_1))}_{\lambda^-(v_0)}\big)$$

I have to show that there is a $n \in \mathbb{N}$, such that $\lambda^-(\mathbf{n})$ is paradoxical, and that for all $m \in \mathbb{N}$, $\lambda^-(\mathbf{m})$ is either false or paradoxical.

It is clear that for all $m \neq gn(\beta)$, $\lambda^-(\mathbf{m})$ is false. Therefore, I only have to show that $\lambda^-(\ulcorner\beta\urcorner)$ is paradoxical:

$$\begin{gathered}\mathcal{V}(\lambda^-(\ulcorner\beta\urcorner) = \mathfrak{p}) \iff \\ \mathcal{V}\big(\ulcorner\beta\urcorner \doteq \ulcorner\beta\urcorner \wedge \exists v_1(\mathbf{Diag}(\ulcorner\beta\urcorner, v_1) \wedge \neg T(v_1))\big) = \mathfrak{p} \iff \\ \mathcal{V}\big(\exists v_1(\underbrace{\mathbf{Diag}(\ulcorner\beta\urcorner, v_1) \wedge \neg T(v_1)}_{\lambda^{--}(v_1)})\big) = \mathfrak{p}\end{gathered} \tag{11}$$

(11) is easily established. To begin with, for any $m \neq gn(\lambda)$, $\lambda^{--}(\mathbf{m})$ is false, since

[30] A word of warning: I certainly do not mean to suggest that Kripke, in this context, would have defined 'paradoxical' as he did in the *Outline*.

$\mathcal{V}(\textbf{Diag}(\ulcorner\beta\urcorner, \textbf{m})) = 0$, for all $m \neq gn(\lambda)$. Furthermore, $\lambda^{--}(\ulcorner\lambda\urcorner)$, i.e.

$$\textbf{Diag}(\ulcorner\beta\urcorner, \ulcorner\lambda\urcorner) \wedge \neg T\ulcorner\lambda\urcorner \tag{12}$$

is paradoxical, since $\mathcal{V}(\neg T\ulcorner\lambda\urcorner) = \mathfrak{p}$ and $\mathcal{V}(\textbf{Diag}(\ulcorner\beta\urcorner, \ulcorner\lambda\urcorner)) = 1$. □

I will not show the details for the Truth-teller being undefined, since they are, *mutatis mutandis*, the same.

We can now turn to the main theorem:

Theorem 3.2. (Metalinguistic T-Schema) For all $\varphi \in \mathscr{L}^t_{pa}$, the following holds:

$$\mathcal{V}(\varphi) \;=\; \mathcal{V}(T\ulcorner\varphi\urcorner)$$

Proof. The proof is quite straightforward, although the details are fairly lengthy. Let me give an outline first: as we know, in Mfp every sentence φ has the same truth value as the sentence $T\ulcorner\varphi\urcorner$. Lemma 3.3 proves that a sentence is true (false) in Mfp if, and only if, it has value 1 (0) in $\langle\mathcal{M}, (E, A, X)\rangle$. This gives us the so-called *Nec* (from φ infer $T\ulcorner\varphi\urcorner$) and *Conec* (from $T\ulcorner\varphi\urcorner$ infer φ): a sentence φ has truth value 1 (0) in $\langle\mathcal{M}, (E, A, X)\rangle$ if, and only if, the sentence $T\ulcorner\varphi\urcorner$ has value 1 (0) too. To complete the proof, it remains to be shown that a sentence φ is paradoxical if, and only if, the sentence $T\ulcorner\varphi\urcorner$ is paradoxical as well. This will be done in Lemma 3.5.

Lemma 3.3. For all $\varphi \in \mathscr{L}^t_{pa}$, the following holds:

$$\langle\mathcal{M}, (E_\infty, A_\infty)\rangle \models_{\text{sk}} \varphi \Leftrightarrow \mathcal{V}(\varphi) = 1$$
$$\langle\mathcal{M}, (E_\infty, A_\infty)\rangle \models_{\text{sk}} \neg\varphi \Leftrightarrow \mathcal{V}(\varphi) = 0$$

Proof. The left-to-right direction

$$\langle\mathcal{M}, (E_\infty, A_\infty)\rangle \models_{\text{sk}} \varphi \Rightarrow \mathcal{V}(\varphi) = 1 \tag{13}$$
$$\langle\mathcal{M}, (E_\infty, A_\infty)\rangle \models_{\text{sk}} \neg\varphi \Rightarrow \mathcal{V}(\varphi) = 0 \tag{14}$$

is evident, since (i) both models have the standard interpretation $\mathcal{M}$ for $\mathscr{L}_{pa}$, (ii) $(E_\infty, A_\infty) = (E, A)$, and (iii) the new logic is exactly like K_3 whenever no conjunct has value $\mathfrak{p}$.

As a shortcut for the right-to-left direction, I will prove that if a sentence has value 1 or 0 in $\langle\mathcal{M}, (E, A, X)\rangle$, then it is not undefined in Mfp. It follows that if a sentence has value 1 (0) in $\langle\mathcal{M}, (E, A, X)\rangle$ then it is true (false) in Mfp, for it cannot be undefined, nor false (true) – otherwise it would have value 0 (1) in $\langle\mathcal{M}, (E, A, X)\rangle$. Let now

'$\langle \mathcal{M}, (E_\infty, A_\infty)\rangle \mathrel{=\!\!\!|}_{\text{sk}} \varphi$' abbreviate '$\varphi$ is undefined in MFP'. It can be shown that

$$(\mathcal{V}(\varphi) = 1 \vee \mathcal{V}(\varphi) = 0) \;\Rightarrow\; \neg\langle \mathcal{M}, (E_\infty, A_\infty)\rangle \mathrel{=\!\!\!|}_{\text{sk}} \varphi$$

A simple induction verifies the statement.

$\boxed{\varphi \equiv T(\mathbf{n})}$ If $\mathcal{V}(T(\mathbf{n})) = 1$ or $\mathcal{V}(T(\mathbf{n})) = 0$, then $n \in E \cup A$ iff $n \in E_\infty \cup A_\infty$ iff $\neg\langle \mathcal{M}, (E_\infty, A_\infty)\rangle \mathrel{=\!\!\!|}_{\text{sk}} T(\mathbf{n})$.

$\boxed{\varphi \equiv \neg\psi}$ If $\mathcal{V}(\neg\psi) = 1$ or $\mathcal{V}(\neg\psi) = 0$, then $\mathcal{V}(\psi) = 0$ or $\mathcal{V}(\psi) = 1$. Thus, by i.h., $\neg\langle \mathcal{M}, (E_\infty, A_\infty)\rangle \mathrel{=\!\!\!|}_{\text{sk}} \psi$ iff $\neg\langle \mathcal{M}, (E_\infty, A_\infty)\rangle \mathrel{=\!\!\!|}_{\text{sk}} \neg\psi$.

$\boxed{\varphi \equiv \psi \vee \chi}$ By contraposition, $\langle \mathcal{M}, (E_\infty, A_\infty)\rangle \mathrel{=\!\!\!|}_{\text{sk}} \psi \vee \chi$ iff at least one disjunct, say ψ, is undefined and the other, say χ, is not true. By i.h., $\mathcal{V}(\psi) \neq 1$ and $\mathcal{V}(\psi) \neq 0$, and therefore $\mathcal{V}(\psi \vee \chi) \neq 0$. To show that $\mathcal{V}(\psi \vee \chi) \neq 1$, it suffices to show that $\mathcal{V}(\chi) \neq 1$, which follows from the fact that χ is either false or undefined in MFP: if it is false, i.e. if $\langle \mathcal{M}, (E_\infty, A_\infty)\rangle \models_{\text{sk}} \neg\chi$ then $\mathcal{V}(\chi) = 0$, and if $\langle \mathcal{M}, (E_\infty, A_\infty)\rangle \mathrel{=\!\!\!|}_{\text{sk}} \chi$, then by i.h. $\mathcal{V}(\chi) \neq 1$. Consequently, $\mathcal{V}(\psi \vee \chi) \neq 1$.

$\boxed{\varphi \equiv \exists v_i(\psi(v_i))}$ By contraposition, $\langle \mathcal{M}, (E_\infty, A_\infty)\rangle \mathrel{=\!\!\!|}_{\text{sk}} \exists v_i(\psi(v_i))$ iff there is no $n \in \mathbb{N}$, such that $\langle \mathcal{M}, (E_\infty, A_\infty)\rangle \models_{\text{sk}} \psi(\mathbf{n})$, and for at least some $n \in \mathbb{N}$, $\psi(\mathbf{n})$ is undefined. Hence, by i.h., for some $n \in \mathbb{N}$, $\mathcal{V}(\psi(\mathbf{n})) \neq 0$, and thus $\mathcal{V}(\exists v_i(\psi(v_i))) \neq 0$. To show that $\mathcal{V}(\exists v_i(\psi(v_i))) \neq 1$, assume the contrary to derive a contradiction. $\mathcal{V}(\exists v_i(\psi(v_i))) = 1$ iff $\exists n \in \mathbb{N}(\mathcal{V}(\psi(\mathbf{n})) = 1)$, iff, by i.h., $\neg\langle \mathcal{M}, (E_\infty, A_\infty)\rangle \mathrel{=\!\!\!|}_{\text{sk}} \psi(\mathbf{n})$. Then either $\langle \mathcal{M}, (E_\infty, A_\infty)\rangle \models_{\text{sk}} \psi(\mathbf{n})$ or $\langle \mathcal{M}, (E_\infty, A_\infty)\rangle \models_{\text{sk}} \neg\psi(\mathbf{n})$. The former implies $\langle \mathcal{M}, (E_\infty, A_\infty)\rangle \models_{\text{sk}} \exists v_i(\psi(v_i))$; the latter implies that $\mathcal{V}(\psi(\mathbf{n})) = 0$, contradicting the assumption. □

LEMMA 3.3 yields the first half of THEOREM 3.2:

Corollary 3.4. (NEC AND CONEC) For all $\varphi \in \mathscr{L}^t_{pa}$, the following holds:

$$\mathcal{V}(\varphi) = 1 \Leftrightarrow \mathcal{V}(T\ulcorner\varphi\urcorner) = 1$$
$$\mathcal{V}(\varphi) = 0 \Leftrightarrow \mathcal{V}(T\ulcorner\varphi\urcorner) = 0$$

Proof. Straightforward consequence of LEMMA 3.3. □

The lemma below completes the proof.

Lemma 3.5. For all $\varphi \in \mathscr{L}^t_{pa}$, the following holds:

$$\mathcal{V}(\varphi) = \mathfrak{p} \;\Leftrightarrow\; \mathcal{V}(T\ulcorner\varphi\urcorner) = \mathfrak{p}$$

Proof. The proof is by induction on the complexity of φ.

$\boxed{\varphi \equiv T(\mathbf{n})}$ $\mathcal{V}(T(\mathbf{n})) = \mathfrak{p}$ iff $n \in X$ iff, by Definition 2.5-(vi), $gn(T(\mathbf{n})) \in X$ iff $\mathcal{V}(T\ulcorner T(\mathbf{n})\urcorner) = \mathfrak{p}$.

Remark 3.6. Notice that we can now use the induction hypothesis $\mathcal{V}(\varphi) = \mathcal{V}(T\ulcorner\varphi\urcorner)$ for all atomic formulas. ■

$\boxed{\varphi \equiv \neg\psi}$ $\mathcal{V}(\neg\psi) = \mathfrak{p}$ iff $\mathcal{V}(\psi) = \mathfrak{p}$ iff, by i.h., $\mathcal{V}(T\ulcorner\psi\urcorner) = \mathfrak{p}$ iff $gn(\psi) \in X$ iff, by Definition 2.5-(ii), $gn(\neg\psi) \in X$ iff $\mathcal{V}(T\ulcorner\neg\psi\urcorner) = \mathfrak{p}$.

Disjunction

$\boxed{\varphi \equiv \psi \vee \chi;\ \Rightarrow}$ $\mathcal{V}(\psi \vee \chi) = \mathfrak{p}$ iff

(A) At least one between ψ and χ, say ψ, is paradoxical.

(B) χ is either false or paradoxical.

From (A),

$$\mathcal{V}(\psi) = \mathfrak{p} \overset{\text{i.h.}}{\Leftrightarrow} \mathcal{V}(T\ulcorner\psi\urcorner) = \mathfrak{p} \Leftrightarrow gn(\psi) \in X \tag{15}$$

Towards a contradiction, assume that $\mathcal{V}(T\ulcorner\psi \vee \chi\urcorner) \neq \mathfrak{p}$, iff

(i) $\mathcal{V}(T\ulcorner\psi \vee \chi\urcorner) = 1$; or

(ii) $\mathcal{V}(T\ulcorner\psi \vee \chi\urcorner) = 0$; or

(iii) $\mathcal{V}(T\ulcorner\psi \vee \chi\urcorner) = \mathfrak{u}$.

We can rule out (i) and (ii), since, by Corollary 3.4, $\mathcal{V}(T\ulcorner\psi \vee \chi\urcorner) = 1(0)$ iff $\mathcal{V}(\psi \vee \chi) = 1(0)$, but we are assuming $\mathcal{V}(\psi \vee \chi) = \mathfrak{p}$. If (iii), then $gn(\psi \vee \chi) \notin X$. On the basis of Definition 2.5-(iii), since we are assuming that $gn(\psi) \in X$, we can argue as follows:

$$(gn(\psi) \in X \Wedge gn(\psi \vee \chi) \notin X) \Rightarrow gn(\chi) \notin A \cup X \tag{16}$$

It follows that either $gn(\chi) \in E$, or $gn(\chi) \notin E \cup A \cup X$. If the former, then $\mathcal{V}(T\ulcorner\chi\urcorner) = 1 \Leftrightarrow \mathcal{V}(\chi) = 1$, and if the latter, then $\mathcal{V}(T\ulcorner\chi\urcorner) = \mathfrak{u} \overset{\text{i.h.}}{\Leftrightarrow} \mathcal{V}(\chi) = \mathfrak{u}$. Both contradict (B). Hence all (i), (ii), and (iii) deliver a contradiction, from which derives that $\mathcal{V}(T\ulcorner\psi \vee \chi\urcorner) = \mathfrak{p}$.

$\boxed{\varphi \equiv \psi \vee \chi;\ \Leftarrow}$ $\mathcal{V}(T\ulcorner\psi \vee \chi\urcorner) = \mathfrak{p}$ iff $gn(\psi \vee \chi) \in X$, iff

(A) At least one between $gn(\psi)$ and $gn(\chi)$, say $gn(\psi)$, is element of X.

(B) $gn(\chi) \in A \cup X$.

From (A)

$$gn(\psi) \in X \;\Leftrightarrow\; \mathcal{V}(T\ulcorner\psi\urcorner) = \mathfrak{p} \overset{\text{i.h.}}{\Leftrightarrow} \mathcal{V}(\psi) = \mathfrak{p} \tag{17}$$

Towards a contradiction, assume $\mathcal{V}(\psi \vee \chi) \neq \mathfrak{p}$. Then – again due to Corollary 3.4 – $\mathcal{V}(\psi \vee \chi) = \mathfrak{u}$. But if $\mathcal{V}(\psi \vee \chi) = \mathfrak{u}$ and $\mathcal{V}(\psi) = \mathfrak{p}$, then $\mathcal{V}(\chi) = \mathfrak{u}$ and therefore, by i.h., $gn(\chi) \notin A \cup X$, which contradicts (B).

Existential Quantifier

$\boxed{\varphi \equiv \exists v_0\psi(v_0);\ \Rightarrow}$ $\quad\mathcal{V}(\exists v_0\psi(v_0)) = \mathfrak{p}$, iff

(A) $\exists n \in \mathbb{N}\,(\mathcal{V}(\psi(\mathbf{n})) = \mathfrak{p})$.

(B) $\forall m \in \mathbb{N}\,(\mathcal{V}(\psi(\mathbf{m})) = \mathfrak{p} \vee \mathcal{V}(\psi(\mathbf{m})) = 0)$.

Using the induction hypothesis, (A) and (B) yield:

(A') $\exists n \in \mathbb{N}\,(gn(\psi(\mathbf{n})) \in X)$.

(B') $\forall m \in \mathbb{N}\,(gn(\psi(\mathbf{m})) \in A \cup X)$.

We derive by Definition 2.5-(v) that $gn(\exists v_0\psi(v_0)) \in X$, and therefore that $\mathcal{V}(T\ulcorner\exists v_0\psi(v_0)\urcorner) = \mathfrak{p}$.

$\boxed{\varphi \equiv \exists v_i\psi(v_i);\ \Leftarrow}$ $\quad\mathcal{V}(T\ulcorner\exists v_i\psi(v_i)\urcorner) = \mathfrak{p}$, iff $gn(\exists v_i\psi(v_i)) \in X$, iff

(A) $\exists n \in \mathbb{N}\,(gn(\psi(\mathbf{n})) \in X)$.

(B) $\forall m \in \mathbb{N}\,(gn(\psi(\mathbf{m})) \in A \cup X)$.

(A) and (B) imply

(A') $\exists n \in \mathbb{N}\big(\mathcal{V}(T\ulcorner\psi(\mathbf{n})\urcorner) = \mathfrak{p}\big)$.

(B') $\forall m \in \mathbb{N}\big(\mathcal{V}(T\ulcorner\psi(\mathbf{m})\urcorner) = \mathfrak{p} \vee \mathcal{V}(T\ulcorner\psi(\mathbf{m})\urcorner) = 0\big)$.

From (A'), we derive by induction that $\exists n \in \mathbb{N}\,(\mathcal{V}(\psi(\mathbf{n})) = \mathfrak{p})$. From (B'), on the other hand, we derive that $\forall m \in \mathbb{N}\,(\mathcal{V}(\psi(\mathbf{m})) = \mathfrak{p} \vee \mathcal{V}(\psi(\mathbf{m})) = 0)$. Therefore, according to the definition of $\mathcal{V}$, $\mathcal{V}(\exists v_i\psi(v_i)) = \mathfrak{p}$. □

Theorem 3.2 derives from Corollary 3.4 and Lemma 3.5. □

4 What's next?

There are two questions I didn't address, which lead to an obvious further step. The first is whether the new logic, together with the new model-theoretical framework, may be useful to deal with other paradoxes. Consider, for instance, the Grelling-Nelson paradox (Grelling and Nelson, 1907) involving the predicate "is heterological"[31]. Within the new framework, one might argue that "'heterological' is heterological" is (like the Liar) paradoxical, for 'heterological' cannot consistently be contained in the extension or in the anti-extension of "is heterological", whereas "'autological' is heterological" is (like the Truth-teller) simply undefined.

The second question is how to obtain a proper *theory* of truth, i.e. how an axiomatisation of the new model may look like[32]. Additionally, one might try to add a "Łukasiewicz conditional" to the new logic, to the effect that $f_{\rightarrow}(\mathfrak{p}, \mathfrak{p}) = 1$. Such a conditional could make $\lambda \leftrightarrow \neg T\ulcorner\lambda\urcorner$ true while both λ and $\neg T\ulcorner\lambda\urcorner$ were still paradoxical. Of course, if one decides to add such a conditional, the interpretation of T must be accordingly modified, in order to preserve the metalinguistic T-Schema. As it is now defined, $gn(\lambda \leftrightarrow \neg T\ulcorner\lambda\urcorner) \notin E$, and hence $\mathcal{V}(T\ulcorner\lambda \leftrightarrow \neg T\ulcorner\lambda\urcorner\urcorner) \neq 1$. Yet, if in the hypothetical new framework $\mathcal{V}(\lambda \leftrightarrow \neg T\ulcorner\lambda\urcorner) = 1$, then its code better be element of E. This seems to me worthy of study[33]: it does seem right to maintain that the Liar sentence is true if and only if untrue. Would it then not be worthwhile to investigate a theory within which both $T\ulcorner\lambda\urcorner$ and $\neg T\ulcorner\lambda\urcorner$ are paradoxical, but where nonetheless $T\ulcorner\lambda\urcorner \leftrightarrow \neg T\ulcorner\lambda\urcorner$ is true?

[31] Mention should be made at this point of the work of Martin, (1967, 1968), who tries to propose one solution for both Liar and Grelling-Nelson paradoxes.

[32] I guess that an appropriate axiomatisation of the model presented here will result in a system somewhere in the neighbourhood of **PKF** (partial Kripke-Feferman).

[33] A study in a similar direction is due to Field, (2002, 2008), who adds a new conditional to K_3, which is not definable as usual by negation and disjunction.

References

Boolos, George S., John P. Burgess, and Richard C. Jeffrey (2007). *Computability and Logic*. 5th edition. Cambridge (UK): Cambridge University Press.

Burgess, John P. (2011). "Kripke on Truth". In: *Saul Kripke*. Ed. by A. Berger. Cambridge (UK): Cambridge University Press, pp. 141–159.

Cantini, Andrea (1989). "Notes on Formal Theories of Truth". In: *Zeitschrift fur mathematische Logik und Grundlagen der Mathematik* 35, pp. 97–130.

Dunn, J. Michael (1969). "Natural Language versus Formal Language". In: *Association for Symbolic Logic*.

— (1976). "Intuitive Semantics for First-Degree Entailments and 'Coupled Trees'". In: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 29.3, pp. 149–168.

Feferman, Solomon (1991). "Reflecting on incompleteness". In: *Journal of Symbolic Logic* 56.1, pp. 1–49.

Field, Hartry (2002). "Saving the Truth Schema from Paradox". In: *Journal of Philosophical Logic* 31.1, pp. 1–27.

— (2008). *Saving Truth from Paradox*. Oxford: Oxford University Press.

Fitting, Melvin (1986). "Notes on the mathematical aspects of Kripke's theory of truth". In: *Notre Dame Journal of Formal Logic* 27.1, pp. 75–88.

Grelling, Kurt and Leonard Nelson (1907). "Bemerkungen Zu den Paradoxien von Russell Und Burali-Forti". In: *Abhandlungen Der Fries'schen Schule (Neue Serie)* 2, pp. 300–334.

Gupta, Anil (1982). "Truth and paradox". In: *Journal of Philosophical Logic* 11.1, pp. 1–60.

Gupta, Anil and Nuel Belnap (1993). *The Revision Theory of Truth*. Cambridge (MA): MIT Press.

Halbach, Volker (2014). *Axiomatic Theories of Truth*. Revised edition. Cambridge (UK): Cambridge University Press.

Halbach, Volker and Leon Horsten (2006). "Axiomatizing Kripke's Theory of Truth". In: *The Journal of Symbolic Logic* 71.4, pp. 677–712.

Halbach, Volker and Albert Visser (2014a). "Self-Reference in Arithmetic I". In: *The Review of Symbolic Logic* 7.4, pp. 671–691.

Halbach, Volker and Albert Visser (2014b). "Self-Reference in Arithmetic II". In: *The Review of Symbolic Logic* 7.4, pp. 692–7121.

Horsten, Leon (2011). *The Tarskian Turn*. Cambridge (MA): MIT Press.

Kleene, Stephen C. (1971). *Introduction to Metamathematics*. Amsterdam: North-Holland.

Kripke, Saul (1975). "Outline of a Theory of Truth". In: *The Journal of Philosophy* 72.19, pp. 690–716.

Martin, Robert L. (1967). "Toward a Solution to the Liar Paradox". In: *The Philosophical Review* 76.3, pp. 279–311.

— (1968). "On Grelling's Paradox". In: *The Philosophical Review* 77.3, pp. 321–331.

McGee, Vann (1991). *Truth, Vagueness, and Paradox: An Essay on the Logic of Truth*. Indianapolis (IN): Hackett Publishing.

Moschovakis, Yiannis N. (1974). *Elementary Induction on Abstract Structures*. Mineola (NY): Dover Publications.

— (2006). *Notes on set Theory*. New York (NY): Springer.

Priest, Graham (1979). "The Logic of Paradox". In: *Journal of Philosophical Logic* 8.1, pp. 219–241.

— (2006). *An Introduction to Non-Classical Logic*. 2nd edition. Cambridge (MA): Cambridge University Press.

Priest, Graham and Francesco Berto (2013). "Dialetheism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Reinhardt, William N. (1986). "Some Remarks on Extending and Interpreting Theories with a Partial Predicate for Truth". In: *Journal of Philosophical Logic* 15.2, pp. 219–251.

Smith, Peter (2013). *An Introduction to Gödel's Theorems*. 2nd edition. Cambridge (UK): Cambridge University Press.

Van Fraassen, Bas C. (1966). "Singular Terms, Truth-Value Gaps, and Free Logic". In: *The Journal of Philosophy* 63.17, pp. 481–495.

Visser, Albert (1984). "Four Valued Semantics and the Liar". In: *Journal of Philosophical Logic* 13.2, pp. 181–212.

Woodruff, Peter W. (1984). "Paradox, Truth and Logic Part I: Paradox and Truth". In: *Journal of Philosophical Logic* 13.2, pp. 213–232.

Yablo, Stephen (1993). “Paradox Without Self–Reference”. In: *Analysis* 53.4, pp. 251–252.