PRINCIPIA **21**(1): 15–21 (2024) doi: 10.5007/1808-1711.2024.e96700
Published by NEL — Epistemology and Logic Research Group, Federal University of Santa Catarina (UFSC), Brazil.

THE MANY FACES OF THE LIAR PARADOX

JOSÉ MARTÍNEZ-FERNÁNDEZ
LOGOS-BIAP/Universitat de Barcelona, SPAIN
jose.martinez@ub.edu

SERGI OMS
LOGOS-BIAP/Universitat de Barcelona, SPAIN
sergi.oms@ub.edu

Abstract. The Liar Paradox is a classic argument that creates a contradiction by reflection on a sentence that attributes falsity to itself: 'this sentence is false'. In our paper we will discuss the ways in which the Liar sentence (and its paradoxical argument) can be represented in first-order logic. The key to the representation is to use first-order logic to model a self-referential language. We will also discuss several related sentences, like the Liar cycles, the empirical versions of the Liar and the Truth teller sentences.

Keywords: Liar paradox • truth • self-reference • semantic paradox

RECEIVED: 15/02/2023 REVISED: 26/04/2023 ACCEPTED: 26/04/2023

1. The Liar paradox

One of the oldest and most venerable paradoxes in the history of philosophy is the Liar Paradox, attributed to Eubulides of Miletus (4^{th} century BC). In its most concise form, it uses the sentence "this sentence is false". Intuitively, if the sentence is true, then what it says is the case, hence it is false. However, if the sentence is false, then what the sentence says is the case, so the sentence is true after all. We find ourselves mired in a paradox.[1]

Alternatively, we could reason to a direct contradiction. Suppose that the sentence is true, then what it says would be the case, hence it would be false. Since it cannot be both true and false, by *reductio ad absurdum* we conclude that our supposition is not true, so the sentence has to be false. However, if the sentence is false, then what the sentence says is the case, so the sentence is true after all. We have proven that the sentence is true and false, a contradiction.

Even though the argument is short and might seem deceptively simple, it has represented a formidable challenge to create a correct theory of truth. For example, a simple solution would postulate that the Liar is neither true nor false. But in that case, let us consider this variation of the paradox: "this sentence is not true". If the

sentence is true, then, as above, the sentence is not true. If the sentence is false, then it is not true, so what the sentence says is the case and it is true. But now we cannot solve the paradox by saying that the sentence is neither true nor false, because if it is neither true nor false, then in particular it is not true, so again what the sentence says is the case and the sentence has to be true. We find ourselves mired in a paradox again.

The situation is even worse because there are a myriad variations of the Liar and a solution should address all of them. For instance, direct self-reference ("this sentence…") is not necessary to construct a paradox. Consider the cyclic Liar: "the following sentence is false", "the previous sentence is true". Both sentences are paradoxical, even though none is directly self-referential. A deeper problem is caused by empirical Liars, i.e., sentences which are paradoxical only under certain empirical conditions. Consider the sentence: "the only sentence written on the blackboard in Room 401 is not true" and let us call it (s). The semantic status of this sentence depends on the facts of the world. If in Room 401 it is only written "the Earth is flat", then (s) is true. If it is only written "$2+2=4$", then (s) is false. But if in that room the only sentence written in the blackboard is (s) itself, then (s) is paradoxical.

In order to explore possible solutions to the paradox it is very important to have a clear understanding of what exactly is involved in all these versions of the paradox. First-order logic, expanded with a truth predicate, gives us the tools to clearly identify the empirical and theoretical presuppositions involved in the Liar arguments. The first step is to create a first-order language that can talk about itself, i.e., that can ascribe properties to its own expressions. We will do that in the next section.

2. Self-referential languages

Let us consider a standard first-order language L built from some collection of individual and predicate constants, denoted $c, d, e, c_1, c_2 \ldots, P, Q, P_1, P_2 \ldots$ respectively. As usual, the language will be interpreted using a structure or model $M = (D, I)$ consisting of a non-empty set D of individuals and an interpretation function I that assigns to each individual constant c a member, $I(c)$, of the domain and to every n-ary predicate constant P a subset, $I(P)$, of D^n. One way of achieving self-reference is simply to include in the domain of interpretation the set of formulas of the language. With the help of individual constants, this already guarantees that the language can achieve self-reference. For instance, let P express the property of being an expression of the language containing the predicate "P". Consider a model in which $I(c) = Pc$, $I(d) = Qc$. Then Pc is true in this model, because $I(c) \in I(P)$ (Pc does contain the symbol P), but Pd is false, because $I(d) \notin I(P)$ (Qc does not contain the symbol P). In this model, Pc formalizes the sentence "this sentence is P". Note that in a different

model in which $I(c) = \neg Pc$, Pc is still true ($\neg Pc$ still contains the symbol P). In this second model, $\neg Pc$ formalizes the sentence "this sentence is not P" which happens to be a false sentence. For the sake of clarity, we will sometimes use the notation $\ulcorner A \urcorner$ to represent any individual constant that gets interpreted as the formula A.[2]

3. Truth in self-referential languages

We now add a new monadic predicate T to the language, which we want to interpret as the truth predicate of the language. Alfred Tarski was the first logician to make explicit that a necessary condition for a predicate T to be the truth predicate for a language L is to satisfy, for every formula A of L, the following biconditional (called a T-biconditional): $T\ulcorner A \urcorner \leftrightarrow A$. The T-biconditionals seem to trivially hold, because to deny them is highly counterintuitive. For instance, let us consider the sentence "snow is white". If we reject its T-biconditional ("snow is white" is true if, and only if, snow is white) we should accept either that "snow is white" is true, even though snow is not white, or that "snow is white" is not true, even though snow is white. But both options seem to be clearly unacceptable. In the proofs of the Liar arguments, we will hence use as axioms all the T-biconditionals for the formulas of the language.

4. The Liar paradoxes formalized

We will use classical logic plus the T-biconditionals that govern the notion of truth. The Liar sentence (in the form "this sentence is not true") can be written as $\neg Tc$, in a model in which $I(c) = \neg Tc$. The T-biconditional of this sentence is: $T\ulcorner \neg Tc \urcorner \leftrightarrow \neg Tc$. Given that $\ulcorner \neg Tc \urcorner = c$, substitution of identicals in the T-biconditional gives us $Tc \leftrightarrow \neg Tc$ and this already expresses the very first version of the paradox we introduced in section 1. We can then reproduce the version of the Liar argument that concludes a direct contradiction in this way:

1.	Tc	supposition
2.	$Tc \rightarrow \neg Tc$	T-biconditional (left-to-right)
3.	$\neg Tc$	*modus ponens* 2, 1
4.	$\neg Tc$	reductio ad absurdum 1-3
5.	$\neg Tc \rightarrow Tc$	T-biconditional (right-to-left)
6.	Tc	*modus ponens* 5, 4
7.	$\neg Tc \& Tc$	&-introduction 4, 6

The cyclic Liar can be easily represented with the use of two constants c and d and two formulas $\neg Td$ and Tc such that $I(c) = \neg Td$ and $I(d) = Tc$.

1.	Tc	supposition
2.	$Tc \rightarrow \neg Td$	T-biconditional of c (left-to-right)
3.	$\neg Td$	*modus ponens* 2, 1
4.	$Tc \rightarrow Td$	T-biconditional of d (right-to-left)
5.	$\neg Tc$	*modus tollens* 4, 3
6.	$\neg Tc$	*reductio ad absurdum* 1-5
7.	$Td \rightarrow Tc$	T-biconditional of d (left-to-right)
8.	$\neg Td$	*modus tollens* 7, 6
9.	$\neg Td \rightarrow Tc$	T-biconditional of c (right-to-left)
10.	Tc	*modus ponens* 9, 8
11.	$\neg Tc \& Tc$	&-introduction 6, 10

Let us consider next our version of the empirical Liar: "the only sentence written on the blackboard in Room 401 is not true". We will formalize this sentence as $\forall x(Px \rightarrow \neg Tx)$, that strictly speaking says: everything that has property P is not true. Let us suppose we are in a model in which there is an individual constant d such that $I(d) = \forall x(Px \rightarrow \neg Tx)$. Now we have to give an interpretation for P that captures the relevant empirical property. We want P to be intuitively the property of being written in the blackboard of Room 401. We choose a model in which only the formula d satisfies property P. This model satisfies the formula $\forall x(Px \leftrightarrow x = d)$. So d is the only formula written in the blackboard of Room 401. In this model, the formula d (i.e., $\forall x(Px \rightarrow \neg Tx)$) says that d is not true. Hence it is a good formalization of our example.

Under these assumptions, the empirical Liar argument could be formalized thus:

1.	Td	supposition
2.	$Td \rightarrow \forall x(Px \rightarrow \neg Tx)$	T-biconditional (left-to-right)
3.	$\forall x(Px \rightarrow \neg Tx)$	*modus ponens* 2, 1
4.	$Pd \rightarrow \neg Td$	$\forall$-elimination 3
5.	Pd	empirical assumption[3]
6.	$\neg Td$	*modus ponens* 4, 5
7.	$\neg Td$	*reductio ad absurdum* 1-6
8.	$\forall x(Px \rightarrow \neg Tx) \rightarrow Td$	T-biconditional (right-to-left)
9.	$\neg \forall x(Px \rightarrow \neg Tx)$	*modus tollens* 8, 7
10.	$\exists x(Px \& Tx)$	first-order logic, 9
11.	$Pe \& Te$	supposition
12.	Pe	&-elimination, 11
13.	$\forall x(Px \leftrightarrow x = d)$	empirical assumption
14.	$Pe \leftrightarrow e = d$	$\exists$-elimination, 13
15.	$e = d$	$\leftrightarrow$-elimination, 14, 12
16.	$Pd \& Td$	substitution of identicals 11, 15

17.	$Pd\&Td$	$\exists$-elimination 10, 11-16
18.	Td	&-elimination, 17
19.	$\neg Td\&Td$	&-introduction 7, 18

5. How to solve the Liar

A lot of ingenuity and technical dexterity has been devoted to the solution of the Liar paradoxes. There are two main groups of proposals: the ones that keep classical logic and the ones that propose an alternative logic. Within classical logic no step in the proof is incorrect and the conclusion cannot be accepted, because it is a contradiction, so some T-biconditionals have to be rejected. The standard solution that keeps classical logic is Tarski's orthodox solution, which forbids the construction of sentences like the Liar. Tarksi creates a hierarchy of truth predicates, such that each truth predicate can be applied to formulas containing truth predicates below in the hierarchy, but not to formulas containing itself or truth predicates above in the hierarchy. This hierarchy of languages cannot express the Liar sentences, even though the T-biconditionals for the sentences expressible in the hierarchy are satisfied.[4]

Some authors criticize the restrictions on self-reference imposed by the Tarskian hierarchy and, in order to avoid the paradoxes, reject classical logic and propose alternative logics. Saul Kripke (1975) has developed a very influential solution using three-valued logics that classify the paradoxical sentences as indeterminate in truth value. Graham Priest accepts that some contradictions are true, in particular, he defends that the Liar is actually true and false. In classical logic, by the rule of *ex contradictione quodlibet* (also called explosion), a contradiction implies anything. So he elaborates a non-classical logic in which that rule is rejected.[5]

Semantic paradoxes keep causing perplexity two millennia after their discovery and are nowadays an active area of research which has deepen our understanding of truth and self-reference and has boosted the creation of non-classical logics.

References

Barwise, J.; Etchemendy, J. 1989. *The Liar*. Oxford: Oxford University Press.

Cobreros, P.; Égré, P.; Ripley, D.; Van Rooij, R. 2013. Reaching Transparent Truth. *Mind* **122** (488): 841–66.

Field, H. 2008. *Saving Truth from Paradox*. Oxford: Oxford University Press.

Glanzberg, M. 2004. A Contextual-Hierarchical Approach to Truth and the Liar Paradox. *Journal of Philosophical Logic* **33**: 27–88.

Gupta, A.; Belnap, N. 1993. *The Revision Theory of Truth*. Cambridge: MIT Press.

Kirkham, R. 1992. *Theories of Truth. A Critical Introduction*. Cambridge Mass.: MIT Press.

Kripke, S. 1975. Outline of a Theory of Truth. *Journal of Philosophy* **72**: 690–716.

Oms, S. 2019. The Sorites Paradox in Philosophy of Logic. In: S. Oms and E. Zardini (eds.). *The Sorites Paradox*, pp. 189-206. Cambridge: Cambridge University Press.
Oms, S. 2023. Some Remarks on the Notion of Paradox. *Acta Analytica* **38**: 211–28.
Priest, G. 2006. *In Contradiction*. New York: Oxford University Press.
Quine, W. V. 1966. The Ways of Paradox. In: *Ways of Paradox and Other Essays*. New York: Random House.
Sainsbury, R. M. 2009. *Paradoxes*. United Kingdom: Cambridge University Press.
Simmons, K. 1993. *Universality and the Liar*. Cambridge: Cambridge University Press.
Smith, P. 2013. *An Introduction to Gödel's Theorems*. New York: Cambridge University Press.
Smullyan, R. 1992. *Gödel's Incompleteness Theorems*. New York: Oxford University Press.
Soames, S. 1999. *Understanding Truth*. New York: Oxford University Press.
Sorensen, R. 2003. *A Brief History of the Paradox*. Oxford: Oxford University Press.
Tarski, A. 1956 [1933]. The Concept of Truth in Formalized Languages. In: *Logic, Semantics, Metamathematics*, pp.152–267. Oxford: Clarendon Press.
Tarski, A. 1944. The Semantic Conception of Truth. *Philosophy and Phenomenological Research* **4**: 341–76.
Zardini, E. 2011. Truth without Contra(di)ction. *Review of Symbolic Logic* **4**(4): 498–535.

Notes

[1]Throughout the paper, we will use a somewhat vague notion of paradox, according to which paradoxes are apparently sound arguments that, apparently, lead to a contradiction. Although this characterization is lacking in precision, it suffices for the purposes of this paper. For further discussion on the notion of paradox see Quine (1966), Sorensen (2003), and Oms (2019, 2023).

[2]Self-reference can also be achieved in a more indirect way. As Gödel showed, in a language in which a basic theory of arithmetic can be represented, formulas can be codified with numbers and $\ulcorner A \urcorner$ can be defined as the name in the language of the code of the formula A. In this case, Gödel's diagonal lemma guarantees the existence of self-referential formulas. This lemma shows that for any open formula $A(x)$, there is a formula B such that the biconditional $B \leftrightarrow A(\ulcorner B \urcorner)$ holds. So B expresses the intuitive sentence "this sentence is A". For more on formalized truth theory, see Smith (2013) and Smullyan (1992).

[3]Notice that this formula follows from the empirical assumption $\forall x(Px \leftrightarrow x = d)$ given the logical truth $d = d$.

[4]Tarski's solution is presented in simplified form in Tarski (1944) and in full in Tarski (1956 [1933]). A good philosophical exposition of Tarski's theory of truth and his solution to the Liar is given in chapters 4, 5 and 9 of Kirkham (1992).

[5]There are also contextualist, revisionist and substructural solutions, which are based on different semantic analyses of the paradoxes. For a short introduction to the Liar and other paradoxes, see Sainsbury (2009). For an accessible and detailed philosophical discussion of Tarksi and Kripke, see Soames (1999). More advanced works include Priest (2006); on revision theory, Gupta and Belnap (1993); on developments of Kripke's theory, Field (2008); on forms of contextualism, Barwise and Etchemendy (1989), Simmons (1993) and Glanzberg (2004); on substructural solutions, Zardini (2011) and Cobreros et al. (2013).

Acknowledgments

We are thankful to the editor of this volume, David Suárez-Rivero, for his helpful comments. The research for this paper has been supported by grant CEX2021-001169-M (funded by MCIN/AEI/10.13039/501100011033); Research Project *Worlds and Truth Values: Challenges to Formal Semantics* (2019PID-107667GB-I00), from the Spanish Ministerio de Ciencia, Innovación y Universidades and Research Project *Unstable Metaphysics* from Ayudas Fundación BBVA a Proyectos de Investigación Científica 2021.