Philosophical Studies (2023) 180:1441–1466
https://doi.org/10.1007/s11098-022-01885-4

A new defense of Tarski's solution to the liar paradox

Gila Sher[1]

Accepted: 3 September 2022 / Published online: 20 November 2022

Abstract
Tarski's hierarchical solution to the Liar paradox is widely viewed as ad hoc. In this paper I show that, on the contrary, Tarski's solution is justified by a sound philosophical principle that concerns the inner structure of truth. This principle provides a common philosophical basis to a number of solutions to the Liar paradox, including Tarski's and Kripke's. Tarski himself may not have been aware of this principle, but by providing a philosophical basis to his hierarchical solution to the paradox, it undermines the ad-hocness objection to this solution. Indeed, it contributes to the defense of Tarski's theory against other objections as well.

Keywords Liar paradox · Tarski · Kripke · Fundamental principle of truth

Tarski's hierarchical solution to the Liar paradox (1933) is widely viewed as ad hoc. This in contrast to non-ad-hoc solutions such as Kripke's (1975). In this paper I show that, on the contrary, Tarski's solution is justified by a sound philosophical principle, which I shall call the "fundamental principle of truth". When we examine truth itself in depth, independently of the paradox, we realize that it has a certain inner structure described by this principle, and this inner structure blocks the classical Liar as well as other related versions. This principle, which is material rather than formal, can be given a variety of formal renditions, and Tarski's hierarchical treatment of the paradox is one of these. Kripke's treatment is another, and for that reason the fundamental principle of truth provides a philosophical basis for Kripke's solution too. Tarski himself may not have been aware of this principle, but it supports, and explains, his solution in a way that undermines the ad-hocness objection. Indeed, it contributes to the defense of Tarski's theory against some other objections as well.

✉ Gila Sher
gsher@ucsd.edu

[1] UCSD, La Jolla, USA

1 Paradoxes in formal and material theories

A paradox in any theory signals a severe error. If the theory is formal in the sense that its subject-matter and/or main methods are logical or mathematical, both the source and the solution to the paradox are likely to be formal. This is the case with logical and set-theoretic paradoxes such as Russell's. But when a paradox emerges in a *material* (non-formal) theory, it is in principle possible that its source is material. An error in understanding the material subject-matter of the theory, or a failure to recognize one of its material principles, may have formal ramifications that generate a paradox. In this case, the key to avoiding paradox may be material, based on a proper understanding of the theory's (material) subject-matter.[1]

The present paper offers a material solution to the Liar paradox. It shows that material considerations lead us to update the equivalence schema in a way that prevents the paradox from arising in the first place. In this it differs from other solutions to the Liar Paradox. While the theory of truth is largely material (rather than formal), philosophers commonly deploy a formal strategy to defend it against paradox. They develop a formal framework for the theory of truth that blocks paradoxes, such as Tarski's (1933) bivalent hierarchical framework or Kripke's (1975) trivalent, externally non-hierarchical, framework.[2]

It is interesting to note that in spite of the fact that philosophers' solutions to the truth paradoxes are formal, their criticisms of Tarski's solution are commonly material. Almost everyone agrees that Tarski's solution is formally powerful: it blocks not just the Classical Liar, but also the Revenge Liar, as well as other versions of the Liar and other semantic paradoxes. But many philosophers criticize Tarski's solution all the same, on material grounds: Tarski's solution is philosophically ad hoc, it does injustice to natural language, it creates many truth predicates instead of one,[3] it relativizes the general concept of truth to language, and so on.

Below I explore the possibility that the Liar paradox is due to our failure to recognize a central material principle of truth. More specifically, I explore the possibility that a certain material principle of truth blocks the paradox, that when we use the truth predicate in accordance with this principle, the truth paradox does not arise.

[1] The distinction between the formal and the material is an old distinction that was understood in many ways through the ages. For the purpose of the present paper, it is sufficient to use logic and mathematics as paradigms of formality and physics, biology, psychology, and most branches of philosophy as paradigms of materiality. Paradoxes are generally formal. In particular, the paradoxes of set theory and the semantic paradoxes are formal. They are logically inconsistent. The theory of truth itself, however, as a theory of, say, the correspondence, coherence, or pragmatic character of truth, is a material theory. Similarly, the current deflationist–substantivist debate is material. When a solution to the Liar is largely guided by considerations that focus on the removal of logical inconsistency—something that holds for virtually all the existent solutions—it is a formal solution; when it is largely guided by general considerations concerning the content or nature of truth, including considerations that focus on the function of truth in the pursuit of knowledge or have to do with the cognitive-epistemic conditions for the possibility of having a full-fledged (correspondence, or coherence, or …) concept of truth, it is material.

[2] My reason for using the "externally" qualifier will become clear in Section 5.2, where we will see that Kripke's solution involves an "internal" hierarchy.

[3] Although this particular criticism is quantitative, it is commonly based on a material consideration: theories of truth should not deviate from natural language, which has only one truth predicate.

This is an unusual way of approaching the paradox, but it brings the theory of truth in line with most other material theories. Normally, in constructing material theories—philosophical or scientific—we are first concerned with their material content and only later check their formal adequacy. But this is not the way philosophers commonly approach the theory of truth. First, they worry about formal adequacy, setting a formal framework that will prevent potential paradoxes from arising, and only later do they set out to develop a correct, comprehensive, and explanatory theory of those aspects of truth they are interested in. This was the order in which Tarski developed his theory of truth: his definition of truth was constructed only after the formal solution to the paradoxes was in place. But this way of developing theories is anomalous in philosophy and science. The strategy explored here treats the theory of truth as a normal material theory.

2 Material theories are not conjunctions

Which material principle of truth blocks the Liar paradox?—Clearly, not the familiar equivalence principle:

Equivalence principle (E)

$$<S> \textit{ is true iff } S,$$

where S is a declarative sentence and "<S>" stands for its name.[4]

In fact, this principle is naturally thought to play a central role in *generating* the paradox. Let us focus on the *classical Liar*.

Liar paradox (LP):

Consider the sentence

(L) L is false.

It follows from E (the Equivalence Principle) plus some widely accepted background principles, such as bivalence and the principles of elementary logic,[5] that:

L is true iff L is false.

Now, it may appear that we cannot block the paradox simply by adding a new material principle to a theory of truth that contains E. Let M be any material principle other than E. It follows from the truth-conditions of the conjunction connective (&) that:

[4] Taking truth-bearers to be sentences is not essential for the present discussion. The discussion can be reformulated in terms of other truth-bearers: propositions, beliefs, thoughts, cognitions, etc.

[5] For the sake of simplicity I assume bivalence here, but what I say can be easily adjusted to non-bivalent discourse.

if

E implies LP,

Then:

E&M implies LP.

Logic teaches us that if we construct our theory of truth as a conjunction of E and M, then regardless of what M is, whatever follows from E also follows from E&M. We cannot save the theory of truth from logical inconsistency by adding a conjunct to E.

Under this description, constructing a material theory is like constructing a conjunction. But generally, constructing a material theory is very different from constructing a conjunction. When we construct a conjunction, adding a conjunct to a sentence S does not change S. But in constructing a material theory—in particular, a philosophical theory—adding another principle, P_2, to a given principle, P_1, often *changes* or *updates* P_1.[6] The addition of principles to a philosophical theory is, generally, quite different from the addition of conjuncts to a sentence:

Updating versus adding a conjunct (★ vs. &)

If P_1 implies S, then $P_1\&P_2$ implies S.

But:

It need not the case that if P_1 implies S, then $P_1 \bigstar P_2$ implies S.

Because:

$P_1 \bigstar P_2$ is equivalent to $P_1{}^*\&P_2$, where $P_1{}^*$ is an updated version of P_1, i.e., $P_1{}^* \neq P_1$.

Accordingly, there can be a material principle of truth, M, such that:

E implies LP,

but

E★M does not imply LP.

Is there such an M?

My answer to this question is positive. The principle I have in mind is the so-called *fundamental principle of truth*. This principle is arrived at by investigating truth beyond the familiar principles, including E.

[6] Updating can take various forms. The form it will take here is restriction: restricting the scope of P_1.

3 The fundamental principle of truth

One way to arrive at the fundamental principle of truth is to begin with a semi-Kantian question: *Under what cognitive conditions can (does) the concept of truth emerge in humans' cognitive life?* What modes of thought do humans have to possess for a full-fledged concept of truth to arise?

To answer this question, we need to place it in an appropriate context. One appropriate context is *cognitive-epistemic*. Starting with the observation that (for one reason or another) it is a central characteristic of human life/culture that we aim to know and understand the world as it is (both practically and theoretically), we further observe that this aim is significantly frustrated by a slew of cognitive limitations (alternatively, by the complexity of the world relative to our cognitive capacities). Given these limitations, we cannot take it for granted that our theories of any aspect of the world are correct. To deal with this cognitive-epistemic predicament, we need a concept that centers on correctness and error. The function of this concept is to address the gap between (a) what our theories (we) say about the world, and (b) how the world is. The concept of truth is assigned the task of addressing this gap. Its function is to distinguish between sentences that get the world right and sentences that do not.[7] To perform this function, the concept of truth must include three parameters:

(i) W—the world as the target of our sentences (theories).
(ii) S—sentences about the world.
(iii) J—judgments about the correctness of S with respect to W.

In this context the question "Under what cognitive conditions can/does truth emerge in human life?" has a clear meaning, a meaning that throws considerable light on the inner structure of the concept of truth.

The answer to this question is given by a fundamental cognitive principle:

***The fundamental principle of truth* (Sher, 2004, 2016)**[8]

Truth requires three basic modes of human thought: an immanent mode, a transcendent mode, and a normative mode.

Let me explain:

(A) *Immanence*. For a significant concept of truth to arise in our life, there must be something that this concept applies to. In the literature, such an object is often called a "truth bearer". In the present cognitive-epistemic context, truth-bearers include sentences and theories, and more broadly, thoughts. But not just any mental or linguistic event is a thought in the sense of a truth-bearer. To be a truth-bearer, a thought has to be directed at something—its "target". It must say something about

[7] In some cases what is right and wrong about the world is sensitive to context, but while the view of truth presented in this paper can, in principle, take this into account, in this paper I will put this issue aside for the sake of simplicity.

[8] In Sher (2004) this principle is called "the immanence thesis".

its target, say something that can be correct or incorrect—true or false—about it. And the thing such a thought must be directed at in order to be true or false is the world (broadly construed), or something in the world, or some facet of the world. So for truth to arise in our cognitive-epistemic life we have to be able to say, or think, something about the world. Without such thoughts, there is nothing about which we can significantly say that it is true or not true, and the concept of truth cannot play its epistemic function in our cognitive life. I call a thought that is directed at the world, and says something about the world or attributes a property (relation) to something in the world, an "immanent thought",[9] and the mode of thought we must have in order to have immanent thoughts, the "immanent mode of thought". *The immanent mode of thought is the mode of directing our mental gaze at some facet of the world and saying something about it, or directing our mental gaze at some objects in the world and attributing some property/relation to them.* A very simple example of an immanent thought is given by the sentence

(1) Snow is white.

This sentence attributes a certain property to something in the world—snow. It attributes to it a property—being white—which it either has or does not have.[10] But immanent sentences are not limited to simple (short, easy to understand) sentences. Many immanent sentences are highly complex.

[9] My use of "immanence" is similar to one of Quine's uses of the term. In some of his writings (e.g., 1981) Quine says that to speak immanently is to speak from *within a theory*, where speaking from within a theory is, typically, saying something about things *outside the theory*, things *in the world* (as distinct from the theory). In my own use, speaking immanently is speaking *in the way* one typically speaks when one speaks from within a theory, namely, speaking about some subject matter, attributing properties/relations to some objects, or saying how the world is. Quine's use highlights a significant dialectic of "immanence": "immanent" connotes "being *internal* to a theory", but "being internal to a theory" signifies "being directed at something *external* to the theory". My idea of immanence is also connected to a common conception of *intentionality* or *aboutness*, expressed in such characterizations as:

> [Intentionality] is that aspect of mental states or events that consists in their being *of* or *about* things (as pertains to the questions, 'What are you thinking of?' and 'What are you thinking about?'). Intentionality is the *aboutness* or *directedness* of mind (or states of mind) to things, objects, states of affairs, events. [Siewert 2002/6: 4].

It is also related to Brentano's conception of intentionality as "reference to a content, direction towards an object …, or immanent objectivity". (Brentano 1995: 88).

My use of "immanence", however, is different from other uses of this term in the philosophical literature, including some of its uses by the authors mentioned above. Thus, in various places Quine characterizes immanent statements as restricted to our mother tongue, a given object language, scientific discourse, or naturalistically construed discourse (see, e.g., Quine 1970/86, 1986, 1995). My own conception of immanence does not impose any of these restrictions. Immanent thought, on my conception, is commonly trans-linguistic: the principles of general relativity, for example, are immanent in my sense, yet they do not belong to a specific language. Not all thoughts, however, are immanent. Thoughts expressing questions, commands, wishes, are not; the Liar sentences we will discuss below are not; and so on.

[10] For simplicity I talk in terms of properties, but what I say can be reformulated in other terms as well. (No specific ontological commitment to properties is required.).

The immanence requirement is a "friction" requirement[11] on truth-bearers. To be a truth-bearer a sentence must have a significant "hook" in the world, be "anchored" in the world. Not every grammatically well-formed sentence satisfies this requirement. Non-immanent sentences are not truth-bearers.

Immanence by itself, however, is not sufficient for truth. To arrive at truth, we need to have, and employ, another mode of thought:

(B) *Transcendence.* To arrive at truth, we have to transcend our immanent thoughts to a standpoint from which we can view not just the world, or that part of the world that our immanent thoughts are directed at, but also our immanent thoughts themselves. I call this mode of thought "the transcendent mode of thought".[12] This mode of thought is universal in the sense that for any immanent thought, we are capable of transcending it to a point of view from which we can see it and say something about it. Specifically, to say that our thoughts about the world are true or false, we need to operate in a special transcendent mode, a mode of thought in which we do not only hold our immanent thoughts in view, but we can in principle evaluate their *correctness*. To operate in this mode, we cannot let go of the world. To hold only our immanent thoughts in view is sufficient for saying, for example, that a given immanent thought is short or long, eloquent or awkward. But to say that a given immanent thought is true or false we need to be able to examine it in relation to the world, or that part of the world it is directed at. Accordingly, for the concept of truth to arise in our cognitive life we must be able to assume a transcendent viewpoint of a special kind, one from which we see both (i) our immanent thoughts and (ii) the world (that part of the world) they are directed at. By a "transcendent mode of thought" I thus understand a mode of thought that enables us to look at both. And *by a "transcendent thought" I mean a thought that has in view both the world and thoughts that say something about the world, attribute properties to things in the world, and so on.* A transcendent thought says something about immanent thoughts in relation to the facet of the world they are directed at. A simple example of a transcendent thought is given by the sentence

(2) "Snow is white" is true.

(2) is clearly transcendent, attributing a property to another sentence, (1), which it holds in view. But (2) is also immanent, because (1) itself is an object in the world, and (2) says something about this object, namely that it is true—that it attributes to some object in the world (snow) a property (being white) that this object has in the world.

Before going on I would like to clarify that by a "transcendent standpoint" I do *not* mean a *Godly* standpoint or "a God's eye view". The transcendence I am talking about is a *human* transcendence: transcending one human standpoint to another human standpoint. Since God, if one exists, does not have cognitive limitations,

[11] For a discussion of friction, see Sher (2016).

[12] This view of transcendence captures the idea of "going beyond" which is commonly associated with transcendence, but it differs from various philosophical terms of art called "transcendence", e.g., Kantian and Husserlian "transcendence". Furthermore, the present notion of transcendence is adjusted to the field of truth: what we transcend (go beyond) is our immanent standpoint, and our transcendent standpoint provides a view of both (i) our immanent thoughts and (ii) the world (or that part of the world) they are directed at.

there is no need for a Godly concept of truth. Truth, as it is considered here, is a *human* concept, and the transcendence it requires is *human* transcendence. We may call such transcendence "H–H transcendence"—human–human transcendence—and distinguish it from “H–G transcendence”—human–God transcendence, i.e., transcendence from a human standpoint to a Godly standpoint. Like Putnam, I believe there is no H–G transcendence. But I also believe that H–G transcendence is not needed for a full-fledged concept of truth to emerge in our cognitive-epistemic life.

I should add that transcendence is a cognitive capacity central to our life not just in connection with truth. Whole disciplines are transcendent in character, for example, the sociology and philosophy of science. Self-reflection—reflection on our thoughts, feelings, actions, etc.—is normally transcendent, both in the moral domain and elsewhere. And so on. These observations highlight the fact that there are multiple transcendent standpoints for a given thought (action, etc.), and multiple things that a given transcendent standpoint enables us to do. Here, however, I focus on truth-transcendence, and more specifically, truth transcendence of *immanent* thoughts.

It is important to recognize that immanence and transcendence are not pairwise disjoint. Many transcendent sentences, such as (2), are immanent. (2) is both immanent and transcendent. (2) is immanent because (i) it is directed at an object in the world, namely, the sentence (1), (ii) it attributes to this object a property that it has or does not have in the world, namely, the property of being true, and (iii) it is itself significantly anchored in the world, through sentence (1). (2) is transcendent because it is directed at a linguistic entity, the immanent sentence (1), and it does so in the right (proper) way: it says something about (1) that has to do with (1)’s connection to the world, and it is itself connected to the world through (1).

We will call sentences that are immanent but not transcendent "*basic* immanent sentences". (1) is a basic immanent sentence, (2) is not. Because for each immanent sentence there is a transcendent truth-sentence, there are infinite chains of transcendent truth-sentences, ultimately connected to basic immanent sentences. I.e., the transcendence ordering is infinite.

The combination of immanence and transcendence, as we have just seen, is central to truth, but it is still not sufficient for truth. To arrive at truth we need a third mode of thought: normativity.

(C) *Normativity.* From a transcendent standpoint we can ask many questions about our theories in relation to the world: whether they target a biological or a psychological facet of the world, whether they use expressions that imitate the sounds of their targets in the world (onomatopoeia), etc. These questions are not questions of truth. Questions of truth are about *correctness*, and in this sense they are *normative*. The property of truth, like the property of justice, is a normative property. A given act or policy has the property of being *just* if it satisfies the *norm of justice*. A given immanent sentence has the property of being *true* if it satisfies the *norm of truth*.

What is the norm of truth? Viewed from our cognitive-epistemic perspective, the norm of truth says that we (our theories) should aim to get the world right. We should not say that snow is white if snow is red. We (our theories) should attribute to objects in the world (or to the world, or to facets of the world) properties and relations they have rather than properties and relations they do not have. When a given sentence satisfies this norm, it has the property of being true. When it does not, it has the property of being false. The norm and property of truth are thus, broadly speaking, a *correspondence* norm and property—not in the naive, simplistic sense of being a copy, picture, or mirror-image of the world, or even of being directly isomorphic with the world, but in the sense of getting the world right by using one or another pattern of correspondence with it, be it direct or indirect, simple or complex.[13]

The notions of immanence, transcendence, and normativity can be extended to predicates and properties. A property attributed to objects by an immanent/transcendent/normative sentence is said to be immanent/transcendent/normative, and a predicate that denotes an immanent/transcendent/normative property is said to be immanent/transcendent/normative as well. The fundamental principle of truth says that truth is a normative transcendent property of immanent thoughts, associated with a (broadly construed) correspondence norm of truth, and that to have a full-fledged concept of truth (a concept which denotes this property), we need immanent, transcendent, and normative modes of thought.

This principle, as we have seen above, sets a constraint on truth-bearers. Using the above-mentioned notion of a basic immanent sentence, we can express this constraint by saying that a truth-bearer must be significantly hooked in the world, either by being a basic immanent sentence or through such a sentence (or sentences). It also sets a constraint on truth-sentences: a truth-sentence is a *proper* truth-sentence iff it is immanent, transcendent, and normative, in the way explained above. Only proper truth-sentences are genuine truth-sentences.

It is easy to see that the fundamental principle of truth has significant ramifications for the equivalence principle E. Specifically, the fundamental principle of truth *updates E* by restricting its scope to *proper* truth-sentences. Let us call the updated equivalence principle "E*". We can now distinguish between the original equivalence principle,

(E) <S> is true iff S, where S is a truth-sentence (proper or improper),

and the updated equivalence principle,

(E*) <S> is true iff S, where "<S> is true" is a *proper* truth-sentence.

4 How the fundamental principle of truth blocks the liar

By updating E in the way described above, the fundamental principle of truth blocks the classical Liar Paradox. We saw how E gives rise to the *Classical Liar*: it follows from E that.

[13] For an example of non-naive correspondence, see, e.g., Sher (2016, 8.4).

(L) L is false.

is true iff it is false. But once E is replaced by E*, the paradox does not arise. E*, unlike E, does not apply to L. To apply to L, L must be a proper truth-sentence, but L is not a proper truth-sentence. L violates all three requirements on a proper truth-sentence: to be transcendent, it must have a standpoint outside the sentence it attributes falsehood to, and to be immanent and (truth-) normative it must be hooked to the world in a significant way, e.g., through the sentence it attributes falsehood to being a basic immanent sentence (or being hooked in the world through such a sentence).

What is new in this way of blocking the Liar paradox is the *grounds* for limiting the E-schema. We limit it not on technical grounds (as a technical means of avoiding the paradox), but on material, philosophical grounds centered on a philosophical understanding of truth, captured by the fundamental principle.[14] These grounds require a revision or an update of E, regardless of whether E leads to a paradox. The paradox is blocked not by introducing a special device that is designed to block it, but by making E materially, or philosophically, sound. As a result, the Liar does not arise in the first place (so there is no need for a special technical device for blocking it).[15]

Does this update of E also result in avoidance of non-classical forms of the Liar? Let us examine a few of these.

Pair Liar.

(3) (4) is false.
(4) (3) is true.

Paradox: If we assume E, it follows that (4) is true iff (3) is true iff (4) is false. This version of the paradox is also blocked by the fundamental principle, since neither (3) nor (4) is a proper truth sentence. These sentences are not truth-normative or immanent for the same reason that L is not, and they fail the transcendence requirement since the transcendence of each is undermined by the other.

Contingent Liar (Kripke)

Suppose Jones made only one statement about Watergate,

[14] This solution is not guided by Occam Razor considerations, which suggest that we make do with a one-pronged ground rather than a three-pronged one. The goal is to block the paradox based on a genuine philosophical understanding of truth, and this requires a three-pronged rather than a one-pronged principle. For the same reason, our ground differs from the mere claim that any sentence requires reference to the world. We ground our solution in the cognitive-epistemic account of the basic human situation, which leads to an explanation of the function of truth in the pursuit of knowledge, which, in turn, leads to the three-pronged fundamental principle of truth, all three components of which are violated by the Liar sentence. In so doing, we give a richer explanation of the reason the Liar sentence is not a proper truth sentence than the mere claim that the Liar sentence fails to refer to the world. Furthermore, by including the immanence-transcendence complementarity as part of the solution, we set the ground for explaining the need for a hierarchical account of truth.

[15] Note that this solution diverges from contextualist solutions which say that the Liar sentence has an "unstable semantic status, switching from defective to non-defective" (Beall, Glanzberg, and Ripley 2011/16). Note also that in spite of this divergence, there is no conflict here: contextualists approach the Liar paradox from a different perspective than that of the fundamental principle, from a natural-linguistic perspective rather than from a cognitive-epistemic perspective.

(5) Most of Nixon's assertions about Watergate are false,

Nixon made the statement,

(6) Everything Jones says about Watergate is true,

and aside from (6), half of Nixon's statements about Watergate are true and half are false.

Paradox: If we accept E, then a paradox ensues: (5) is true iff (6) is false iff (5) is false. But if we replace E by E*, the paradox is allayed. Under the given circumstances, neither (5) nor (6) is a proper truth-sentence, for the same reason that neither (3) nor (4) is a proper truth sentence: under the given circumstances, what (5) and (6) say come down to what (3) and (4) do.

Infinitely Descending Liar (Yablo)

This version of the Liar consists of an infinite descending chain:

(S_1) For all $k > 1$, S_k is false
(S_2) For all $k > 2$, S_k is false
(S_3) For all $k > 3$, S_k is false

⋮

Paradox: Either (a) some S_n is true or (b) all the S_n's are false. (a): Suppose S_n is true. Given E, S_{n+1} is false. Then for some $k > n + 1$, S_k is true. Then S_n is false. Contradiction. (b): Suppose all the S_n's are false. Given E, all the S_n's are true. Contradiction.

To examine this paradox in light of the fundamental principle of truth, we have to allow infinite chains of truth-sentences. This by itself need not violate the immanence, transcendence, or normativity requirement. So we can go on. But a descending infinite chain like Yablo's is blocked by the fundamental principle since both the requirement of being immanent (significantly hooked in the world, e.g., through a basic immanent sentence) and the requirement of being properly normatively-transcendent (i.e., attributing correctness/incorrectness-in-the-world to an immanent sentence), are violated.

Trivalent Liar (Strengthened Liar, Revenge Liar)

Consider

(7) (7) is not true,

where assuming trivalence, a sentence is not true iff it is either false or indeterminate. Assuming E, a paradox arises: (7) is true iff (7) is not true. [If (7) is false, then (7) is not true, then (7) is true, then (7) is not false. If (7) is indeterminate, then (7) is not true, then (7) is true, then (7) is not indeterminate.]

To evaluate this version of the paradox from the point of view of the fundamental principle of truth we have to allow trivalence. This by itself does not violate any of the immanence, transcendence, and normativity requirements of the fundamental principle, so we can proceed. But (7) is not a proper truth-sentence, for the same reasons that L (the classical Liar) is not.

We have seen that there is at least one material principle of truth that blocks the Liar paradox (in multiple versions): the fundamental principle of truth. It is important to recognize that the fundamental principle of truth is not an ad hoc principle for blocking paradoxes. It is a substantive philosophical principle that identifies a significant (3-pronged) material feature of truth, and it blocks sentences that violate this principle whether they are paradoxical or not.

Consider the sentence

(8) (8) is true.

This sentence is not paradoxical, yet it is an improper truth-sentence for the same reason that L is, and as such it is blocked by the fundamental principle of truth.

5 The fundamental principle of truth as a philosophical basis for Tarski's solution to the liar paradox

We have seen that a material method (centered on the fundamental principle of truth) blocks the Liar paradox. And we know that various formal methods, such as Tarski's and Kripke's, block it as well. Is there a need for both material and formal methods for blocking the paradox? My answer to this question is positive: the formal solutions to the Liar paradox, such as Tarski's and Kripke's, require a substantial philosophical basis, like the one given by the fundamental principle of truth, and the fundamental principle requires formal renditions, such as those provided by Tarski and Kripke, to precisify it.

The fundamental principle of truth is a philosophically substantial principle. But while it is effective in blocking central cases of the Liar paradox, it lacks a detailed presentation that would enable it to make a definite judgment in all possible cases. It is a big-picture principle and it expresses a clear philosophical idea, but just this prevents it from addressing absolutely all cases, including all borderline cases. This situation calls for the development of systematic, possibly *formal, rendition* of this principle. By a "formal rendition of a philosophical idea/principle" I mean a reformulation or a development that provides a clear, sharp, and formally rigorous structure to this idea/principle, one that enables it to cover all cases whatsoever.

But such a formal development cannot replace the fundamental principle. By itself, it may very well lack an adequate philosophical basis. Kripke, if not Tarski, recognized this problem:

> I do not regard any proposal, including the one to be advanced here [of solving the semantic paradoxes], as definitive ... On the contrary, I have not at the

> moment thought through a careful philosophical justification of the proposal. (Kripke, 1975: 699)[16]

The fundamental principle of truth provides a philosophical basis for Tarski's and Kripke's formal solutions to the Liar paradox. It shows that these formal solutions are grounded in a sound and substantial philosophical principle of truth. In a sense, they are formal renditions of this principle.

In showing how these solutions are philosophically grounded in this way, we have to be aware of the fact that formal renditions of material philosophical principles often come with a price. They might distort certain aspects of the original ideas, focus on things that are not central to these ideas, introduce elements that are irrelevant from the perspective of these ideas, burden us with superfluous commitments, and so on. Furthermore, we need to keep in mind that the same idea can be treated formally in different ways, each with its own strengths and weaknesses, that there is rarely a perfect fit between a philosophical idea and its formal rendition, and that the choice of a formal treatment is often based on tangential goals and/or pragmatic considerations. As a result, the same philosophical principle can support competing solutions to a given problem, as is the case with the fundamental principle of truth and different solutions to the Liar paradox, such as Tarski's and Kripke's.

It has been, however, only Tarski's solution which was criticized as being ad-hoc. For that reason, the present defense is primarily, if not exclusively, a defense of his solution.

***1. Tarski's solution to the liar paradox*.** A few significant characteristics of Tarski's formal solution to the Liar paradox (in the present context) are:

A. Tarski's formal solution to the Liar paradox is constrained by his (material) declaration:

> [T]hroughout this work I shall be concerned exclusively with grasping the intentions which are contained in the so-called *classical* conception of truth ('true – corresponding with reality'). (Tarski, 1933: 153)

Tarski's correspondence approach is also reflected in his understanding of semantics:

> We shall understand by semantics the totality of considerations concerning those concepts which, roughly speaking, express certain connexions between the expressions of a language and the objects and states of affairs referred to by these expressions. (Tarski, 1936b: 401)

It is reasonable to view "Convention T" (1933: 187–8), Tarski's material adequacy condition for a definition of truth, as expressing his correspondence conception of truth in the 1933 paper. Granted, Tarski formulated this condition in a way that is similar to a *disquotational* version of the equivalence principle, and this may lead one to view it as a purely linguistic condition. But Tarski believed that

[16] I should note, though, that Kripke's paper includes extended intuitive remarks focused on linguistic observations. But as the above citation indicates, Kripke himself did not regard these intuitive remarks as providing the requisite philosophical grounding.

correspondence can be expressed in a purely linguistic form: "concepts … that … give expression to certain relations between the expressions of language and the objects about which these expressions speak" can be viewed also as "set[ing] up [a] correlation between the names of expressions and the expressions themselves" (*ibid.:* 252). Philosophically, even though the biconditional "'Snow is white' is true iff snow is white" can be viewed as relating one linguistic expression—"'snow is white'"—to another—"snow is white", this does not conflict with its being a correspondence-with-the-world statement: "Snow is white" is true iff in the world, the stuff snow has the property of being white.

The correspondence character of Tarski's conception of truth is reflected, in his hierarchical solution to the Liar paradox, in the fact that the lowest language in any Tarskian hierarchy is a *pure object language*, a language whose sentences speak directly about the world.[17]

B. Tarski's formal solution to the Liar paradox is limited to so-called *open* languages, languages that do not contain their own semantic apparatus. In particular, an open language L does not include the predicates "x is true", "x refers to y" ("x denotes y"), "x is satisfied by y", etc., applied to its own expressions. This excludes natural language, which is *closed* in Tarski's sense. Among open languages, Tarski further limited his attention to "formalized languages of the deductive sciences"—essentially languages formulated within the framework of modern logic.

C. Each formalized language is built by Tarski as an infinite hierarchy of partial languages: L_0, L_1, L_2, … .[18] L_0 is a *[pure] object-language (OL)*. Its non-logical constants denote objects and properties in the world (broadly understood), but it has no predicates that function as semantic predicates. L_1 is the so-called *meta-language (ML) of L_0(OL)*. ML has a logical apparatus that includes, and is at least as strong as, that of OL. Its non-logical vocabulary includes (i) names of all the expressions of OL, (ii) semantic predicates applicable to expressions of OL, including a truth-predicate, T_1, for OL, (iii) non-logical constants that enable ML to talk about the world, including the capacity to say everything that OL can say about it, and possibly (iv) additional non-logical vocabulary. In particular, if S is a sentence of OL, ML has both sentences that assign a truth-value to S and sentences that say whatever S says about the world.

It is important to recognize that although ML is a meta-language, it is also an object language. ML, like OL, has resources for talking about the world—indeed about a larger swath of the world than OL can talk about, since it can talk about linguistic expressions of OL and attribute semantic properties to these expressions. Depending on its vocabulary, it may also have resources for saying things about the world that OL cannot.

ML, however, does not include either names of its own expressions or semantic predicates that apply to its own expressions. To talk about the truth of ML-sentences we have to ascend to MML. MML has a truth-predicate that applies to sentences of ML, T_2. It is related to ML and to the world in a way parallel to that in which ML is

[17] The term "pure", applied to object-languages (here and elsewhere in this paper), is a term that I add to Tarski's terminology. This term, however, is clearly implicit in the Tarskian conception.

[18] The hierarchy extends to transfinite languages (e.g., L_ω), but I will not dwell on the transfinite portion of the hierarchy.

related to OL and the world. The truth predicate of MML—T_3—belongs to MMML, and so on. Each language L_n has a meta-language, L_{n+1}, with a truth-predicate, T_{n+1}, that applies to sentences of L_n. The use of any truth-predicate T_{n+1} is subject to the equivalence principle of L_{n+1}: "T_{n+1}<S_n> iff S_{n+1}", where "S_n" is an L_n sentence, "<S_n>" is an L_{n+1} name of S_n, and "S_{n+1}" is an L_{n+1} sentence that says about the world whatever S_n says about it. Only truth-sentences that accord with the above description are semantically well-formed in Tarski's account.

The classical Liar paradox is blocked by the Tarskian hierarchy. Classical Liar sentences violate the rules of the Tarskian hierarchy and as such do not belong to any Tarskian language L_n. These sentences are not well-formed (either syntactically or semantically), and as such they are banned by Tarski. In addition to the classical Liar, Tarski's hierarchical system blocks all the non-classical Liar sentences mentioned above, as well as all other semantic paradoxes discussed in the literature. But it also blocks non-paradoxical sentences that violate its principles, like (8). It is worth noting that, as we have seen above, the fundamental principle of truth also regards (8) as an improper truth sentence, based on substantive philosophical considerations.

2. How the fundamental principle of truth defends Tarskian solution to the liar paradox against the ad-hocness charge.

The Tarskian solution to the Liar paradox involves a hierarchical conception of language. Tarski himself justified this conception as an effective device for blocking the Liar and other semantic paradoxes. But this hierarchical conception was criticized by many philosophers as ad hoc. Kirkham (1992) formulated this criticism as follows: "Tarski has no independent reason for postulating the distinction between object language and metalanguage other than to solve the Liar Paradox" (*ibid.:* 281). Attempts to refute this criticism are very rare. The present paper presents a refutation of this criticism based on the fundamental principle of truth. Regardless of Tarski's own reasons for his hierarchical conception of language, his conception is grounded in this substantive philosophical principle.

To see how the fundamental principle of truth defends various aspects of the Tarskian hierarchy, consider the following table, which pairs various elements of the fundamental principle with parallel elements of the Tarskian hierarchy, in particular, various types of sentences with various Tarskian languages[19]:

Fundamental Principle	*Tarskian Hierarchy*
Basic immanent sentence	Pure object-language (L_0)
Transcendent truth-sentence	Meta-language (ML)
Every immanent sentence S has a transcendent truth-sentence, "<S> is true/false"	For every language L, there is a meta language ML, with truth-sentences "<S> is true/false" for sentences S of L
If the sentence S is immanent, then the transcendent truth-sentence "<S> is true/false" is also immanent.[20]	If L_n is an object-language, then ML_n is also an object-language

[19] In presenting this table I am primarily concerned with clarity (rather than with economy).

[20] This generalizes straightforwardly to truth-sentences which attribute truth-values to more than one sentence.

The relation "X is transcendent to Y" between sentences requires that X be transcendent to Y in a significant way	The relation "X is a meta-language of Y" between languages is a strong partial-ordering, forming linear chains called "hierarchies"
The transcendence ordering is infinite	The language−meta-language hierarchy is infinite
A proper truth-sentence	A well-formed truth-sentence[21]

Similar parallels hold between the fundamental principle of truth and the conception of language involved in Kripke's solution to the Liar paradox, something which suggests that Tarski's and Kripke's solutions are not as far apart as they may appear to be. Where Tarski's solution invovles an *external* hierarchy of languages, Kripke's solution involves an *internal* hierarchy of stages of determining the extension and anti-extension of the one truth-predicate. And where Tarski's solution involves many truth-predicates, Kripke's one truth-predicate is "built" in multiple stages corresponding to Tarski's multiple truth-predicates. (The parallels are especially strong in the finite levels of the two hierarchies.)

As for the parallels between Kripke's hierarchy and the fundamental principle of truth, the idea of *immanence* is represented in Kripke's hierarchy by the concept of *groundedness* and the idea of *transcendence* by a *stage by stage* determination of the extension (and anti-extension) of the truth predicate. Kripke informally described the intuition underlying his concept of groundedness as follows:

> It has long been recognized that some of the intuitive trouble with Liar sentences is shared with such sentences as [(8)] which, though not paradoxical, yield no determinate truth conditions. ... In general, if a sentence ... asserts that (all, some, most, etc.) of the sentences of a certain class C are true, its truth value can be ascertained if the truth values of the sentences in the class C are ascertained. If some of these sentences themselves involve the notion of truth, their truth value in turn must be ascertained by looking at *other* sentences, and so on. If ultimately this process terminates in sentences not mentioning the concept of truth, so that the truth value of the original statement can be ascertained, we call the original sentence *grounded*; otherwise, ungrounded. ... Sentences such as [(8)], though not paradoxical, are ungrounded. (Kripke, 1975: 693–4)

The grounded sentences in Kripke's system are all immanent in the sense of the fundamental principle.

The idea of transcendence is represented by Kripke's conception of the extension (and anti-extension) of the (one) truth predicate, T, of the universal language, $\mathscr{L}$, as determined *in stages*: stage 0, stage 1, stage 2, etc., which he likened to Tarskian languages: L_0, L_1, L_2, etc. Simplifying, each stage consists of two sets, Σ_1—the extension of T, i.e., the set of all sentences of $\mathscr{L}$ which are determined to be true in this stage, and Σ_2—the anti-extension of T in this stage, i.e., the set of all sentences of $\mathscr{L}$ which are determined to be false in this stage. The determination of truth in the ascending stages is monotonic, i.e., determinations of truth/falsehood in earlier

[21] A sentence that satisfies both the syntactic and the semantic constraints set by Tarski on languages in which, and languages for which, truth is definable.

stages are not changed in later stages. In stage 0 both Σ_1 and Σ_2 are empty. This stage represents the basic immanent sentences of the fundamental principle of truth. Sentence (1) belongs here. In stage 1 the truth and falsehood of the basic immanent sentences is determined. Sentence (2) belongs here. In stage 2 the truth and falsehood of proper truth-sentences whose truth-value is determined in stage 1 is determined. And so on. If S is an ungrounded truth-sentence, S is neither in the extension nor in the anti-extension of the truth predicate in any of the finite stages. Within these stages, Kripke's solution can be viewed as a (particular) precisification of the fundamental principle of truth. In Kripke's system, the stage in which the truth-value of a given sentence is determined can be affected by contingent circumstances. This, too, is compatible with (though not essential for) the fundamental principle. Consider the sentence:

(9) The first two numbered immanent sentences in paper A are true,

where A is some given paper with at least two numbered immanent sentences. (9) is transcendent, but its level of transcendence depends on a contingent fact concerning what the first two numbered immanent sentences in paper A are. If both are basic immanent sentences, the level of transcendence of (9) is 1; otherwise, higher.

Once we get to transfinite levels of the hierarchy, Kripke's solution diverges from the fundamental principle in certain ways. For example, there are levels in which (8) is true, others in which it is false, and still others in which it is indeterminate.[22] According to the fundamental principle, (8) is not a proper truth sentence on any level. Still, there are significant core parallels between the Kripke's solution to the Liar paradox and the fundamental principle of truth.[23]

Indeed, although Tarski's and Kripke's solution differ from each other in a variety of ways, both receive a substantive, material, philosophical grounding by the fundamental principle of truth, and as such are not ad hoc. So does any solution that incorporates some (sufficiently strong) version of the immanence, transcendence, and normativity subprinciples. This points to the universality of the fundamental principle of truth, a principle that in hindsight can be said to be implicit in most approaches to truth but has never before been explicitly stated or recognized.

The ad-hocness charge against Tarski's hierarchical approach to truth is often directed at the very idea of a meta-language. We have grounded this idea philosophically in the transcendence of truth, but it is worth noting that the idea of a

[22] See Kripke (1975: 708–9).

[23] At this point, some readers may wonder in what way the fundamental-principle solution to the Liar paradox differs from Kripke's [aside from the point concerning (8)]. First, Kripke's solution is based only on the groundedness of sentences, which corresponds to our immanence, but our solution is also based on the principles of transcendence and normativity. Second, as we recall from the above citation from p. 699 of Kripke's paper, Kripke did not offer a philosophical justification for his solution, but we offer such a justification for our solution. (Kripke did clarify his notion of grounding by referring to the way we explain "true" to someone who does not yet understand it. But, as we see from the above-mentioned citation, he did not consider this a "careful philosophical justification". Furthermore, our cognitive-epistemic justification is quite different from Kripke's clarification in terms of learning.) Finally, the present solution is more general than Kripke's, being instantiated by both Kripke's and Tarski's solutions.

meta-language has other credentials as well. This idea, and the related idea of a meta-theory, have proven extremely fruitful in a variety of fields. Gödel's completeness and incompleteness theorems, Tarski's definition of "logical consequence", Church's thesis, Turing's proof of the unsolvability of the halting problem, the Löwenheim–Skolem theorem, Lindström's theorems—are all extremely valuable meta-linguistic theorems. In fact, most of what is taught today in non-elementary logic classes is meta-logic, and both model-theory and proof-theory are essentially meta-theoretical. Parts of linguistics—e.g., linguistic semantics—are also meta-theoretical, and so is meta-philosophy.

Furthermore, it is important to recognize that even solutions to the Liar paradox which allegedly demonstrate the dispensability of the language−meta-language duality, in fact appeal to this duality.[24] Kripke is fully cognizant of this point:

> It seems likely that many who have worked on the [non-Tarskian] approach to the semantic paradoxes have hoped for a universal language, one in which everything that can be stated at all can be expressed. ... Now the languages of the present approach contain their own truth predicates and even their own satisfaction predicates, and thus to this extent the hope has been realized. Nevertheless the present approach certainly does not claim to give a universal language, and *I doubt that such a goal can be achieved*. First, the induction defining the minimal fixed point is carried out in a set-theoretic *meta-language*, not in the object language itself. Second, *there are assertions* we can make about the object language *which we cannot make in the object language*. For example, Liar sentences are *not true* in the object language, in the sense that the inductive process never makes them true; but we are precluded from saying this in the object language by our interpretation of negation and the truth predicate. If we think of the minimal fixed point, say under the Kleene valuation, as giving a model of natural language, then the sense in which we can say, in natural language, that a Liar sentence is not true must be thought of as associated with some *later stage* in the development of natural language, one in which speakers *reflect* on the generation process leading to the minimal fixed point. It is not itself a part of that process. The *necessity to ascend to a metalanguage* may be one of the weaknesses of the present theory. *The ghost of the Tarski hierarchy is still with us*. (Kripke, 1975: 714, my italics)

3. Defense of Tarski's theory against additional criticisms.

Other criticisms of Tarski's theory of truth include: (a) relativity of truth to language and multiplicity of truth-predicates, (b) doing injustice to natural language, (c) extreme restrictions, (d) diminished explanatory power, and (e) infinite ascent (regress) and absolute generality. Let me briefly describe these criticisms and offer a defense against them. In some cases the defense is related to the fundamental principle of truth; in others it is independent of it.[25]

(a) *Relativity of truth to language; multiplicity of truth-predicates.*

[24] They do not only have parallels with this duality but actually appeal to it.

[25] I include the latter cases for the sake of completeness.

Criticism: Tarski's theory generates definitions of truth for particular languages: the language of set-theory, the language of Boolean algebra, the language of arithmetic, the language of special relativity, and so on, where each language has its own truth-predicate. In so doing it *relativizes* truth to languages and postulates a *multiplicity* of truth-predicates. In fact, however, there is only *one* truth-predicate, and this predicate expresses an *absolute, non-relativized* notion of truth. Blackburn (1984: 267) compared the procedure of defining *truth* by defining truth-in-L_A, truth-in-L_B, …, for different languages L_A, L_B, …, to defining the notion of *properly grounded verdict* by defining "properly-grounded-verdict-on-Monday", "properly-grounded-verdict-on-Tuesday", …, for different days of the week. Just as there is no philosophical interest in the *relative* jurisprudential notion "properly-grounded-verdict-on-day-X", so there is no philosophical interest in the *relative* semantic notion "true-in-L".

Defense: First, let us get the apparent relativity of Tarski's notion of truth to language straight. There are two kinds of relativity here: (i) Relativity to different languages—L_A, L_B, …—each with a different interpreted non-logical (and non-semantic) vocabulary and its own truth-predicate—T_A, T_B, …. (ii) For each language L_A, L_B, …, relativity to a language in its Tarskian hierarchy of object- and meta-languages: L_0, L_1 (=ML_0), L_2 (=MML_0), … and multiplicity of truth predicates—T_1, T_2, …. (So, overall we have languages L_{A0}, L_{A1}, L_{A2}, …, L_{B0}, L_{B1}, … and truth predicates T_{A1}, T_{A2}, …, T_{B1}, ….)

Now, the attribution of the first type of relativity to Tarski's theory appears to be based on the fact that Tarski presented his definition of truth by selecting a *particular* formalized language, the pure object-language of Boolean algebra (the "calculus of classes"), with its specific non-logical vocabulary, and defining truth for it, rather than by taking an *arbitrary* formalized, pure object-language, with an arbitrary non-logical vocabulary, and defining truth for it. This creates the impression that his definition of truth is specific to a particular language, or that each language has its own, separate, definition of truth. But this impression is incorrect. Tarski explained that he chose to present his definition in this way for a practical reason: ease of understanding the definition. In principle, it is possible to formulate his definition of truth as a general definition applicable to an arbitrary object-language (selection of non-logical constants):

> For an extensive group of formalized languages it is possible to give a *method* by which a correct definition of truth can be constructed *for each of them*. The *general abstract description* of this method and of the languages to which it is applicable would be troublesome and *not at all perspicuous. I prefer therefore* to introduce the reader to this method *in another way*. I shall construct a definition of this kind in connexion with a particular concrete language and show some of its most important consequences. The indications which I shall then give in §4 of this article will, I hope, be sufficient to show how *the method illustrated by this example can be applied to other languages of similar logical construction.* (Tarski, 1933: 167–8. My italics)

But if the definition of truth can in principle be defined for an arbitrary language (arbitrary non-logical vocabulary), then from a philosophical perspective, the first type of relativity, and the related multiplicity of truth-predicates, are merely notational and insignificant.

The second type of relativity and multiplicity of truth predicates—relativity to L_0, L_1, … and truth-predicates T_1, T_2, …—is different from the first. This so-called relativity is more accurately categorized as a hierarchy, or as one form the hierarchy of truth can take, and this particular hierarchy, along with others, is, as we have explained above, justified by the nature of truth itself, as captured by the fundamental principle. The Tarskian hierarchy formally represents two central philosophical features of truth, *immanence* and *transcendence*, using the devices of *basic object language*, *ascent to a meta-language*, and *multiple truth-predicates.* But while it is possible to express the transcendence of truth by a variety of other technical devices, including ones that do not involve an external hierarchy of object- and meta-languages or a multiplicity of truth-predicates, as, e.g., in Kripke's theory, from the philosophical perspective of the fundamental principle, the Tarskian hierarchy is perfectly sound. Indeed, it has some advantages over the Kripkean hierarchy.[26] For example:

(a) The Tarskian hierarchy rejects (8) on all levels, in accordance with the fundamental principle of truth, while Kripke's hierarchy does not.
(b) While the Kripkean hierarchy, as recognized by Kripke,[27] makes essential use of the Tarskian hierarchy at some point, the Tarskian hierarchy does not make any use of the Kripkean hierarchy at any point.
(c) Tarski's hierarchy and its principles block a wider range of semantic paradoxes than many other solutions to the Liar paradox, including Kripke's.

Furthermore, historically, the Tarskian hierarchy proved extremely fruitful in metalogic, including proof-theory and logical semantics (model theory). It is an open question whether the Kripkean hierarchy would have proven equally fruitful.

In this context, it is worthwhile to compare Tarski's (1933) definition of truth to his (1936a) definition of logical consequence. His definition of logical consequence is *technically* just as relative to language in the two ways indicated above: here, too, he defines logical consequence for different languages—languages with different vocabularies—and he assumes a hierarchy of object- and meta-languages. So technically, there appears to be relativity to languages and multiple logical-consequence relations. Yet, as far as I know, there are no criticisms of Tarski's definition of logical consequence on the ground that it renders the notion of logical consequence relative to language or that it involves a multiplicity of notions of logical consequence. And for a good reason.

Why are there no such criticisms? Possibly because Tarski presented his definition of logical consequence more briefly, with fewer details and fewer examples than his definition of truth. Why is there a good reason for not raising such criticisms?

[26] By saying this, I do not say that Kripke's hierarchy has no advantages compared to Tarski's. It does. But due to the fact that (i) Tarski's hierarchy is so often thought to be inferior to Kripke's, (ii) the present paper is devoted to a defense of Tarski, I focus on ths advantages of Tarski's hierarcy.

[27] See citation from p. 714 of his 1975 paper above.

Because the "relativity" of logical consequence to language (and the related multiplicity of logical-consequence predicates) is a mere technical device.

Finally, Blackburn's analogy is misleading. Whereas there is no significant connection between proper legal verdicts and days of the week, there is a significant connection between truth and transcendence.

(b) *Doing injustice to natural language.*

Criticisms: To condense the discussion, let me focus on two criticisms: (i) Tarski claimed that natural language is an inconsistent language since it cannot avoid the Liar paradox; this claim in incorrect/unwarranted. (ii) Tarski's theory of truth fails to accomplish the main task of a theory of truth, namely, define truth for natural language.

Defense: First, let me note that even if criticism (i) is correct, this by itself does not undermine the theoretical usefulness, value, and fruitfulness of Tarski's solution to the Liar paradox, which are independent of the question of natural language. Whether criticism (i) is in fact correct I leave an open question. But I do take seriously Kripke's observation that if we are engaged in a *theoretical* account of the semantic structure of language, we need Tarski's language–meta-language duality (see citation above) or something like it.

Be that as it may, the more relevant criticism is (ii). This criticism raises an important methodological question concerning philosophy: to what extent should philosophical theories of subject-matters such as truth, knowledge, ontology, logical consequence, etc. focus on the use of these notions in natural language? Leaving the adjudication of this general question to another occasion, it is quite clear that the linguistic perspective is not the only worthwhile philosophical perspective and the linguistic task is not the only significant task of philosophy. In particular, the cognitive-epistemic perspective underlying the fundamental principle of truth is no less important than the linguistic perspective. The fact that the Tarskian solution to the Liar paradox is grounded in a philosophically significant (cognitive-epistemic) principle of truth, that it introduces, or is based on, a duality (of language and meta-language) that has proven extremely fruitful in a number of disciplines, that it is extraordinarily efficient in blocking paradoxes, and that the theory/definition of truth given within the parameters of this solution has important applications in logic, mathematics, linguistics, and possibly other disciplines – all these strongly suggest that, independently of the natural-language perspective, Tarski's solution and theory/definition are highly valuable.

(c) *Excessive restrictions.*

Criticism: The Tarskian solution sets excessive restrictions on the use of language, for example, a total ban on self-reference.

Defense: Tarski did not set a *total* ban on self-reference. He did not ban Gödelian self-reference, which uses syntax to refer to syntax in proof-theoretic contexts. The ban is limited to semantic contexts. Furthermore, in general, the building of a highly-efficient, clean, simple, and aesthetically-pleasing system often involves setting rigid restrictions. Tarski's choices appear to reflect his priorities.

(d) *Diminished explanatory power.*

Criticism: Some philosophers claim that Tarski's hierarchical account of truth sets unreasonable limits on our theoretical goals. Thus, McGee says:

> At issue is the possibility of a unified scientific understanding in which human thought and action are no less intelligible or more mysterious than the planetary orbits. If we adopt [Tarski's] proposed solution, we shall find that within the object language we are unable even to describe human thought and action. ... [I]ntentional human activities, such as speaking, believing, ... will be indescribable and inexplicable. Within the metalanguage we can obtain fragmentary descriptions of human thought and actions. ... [B]ut thought about thought and talk about talk will remain indescribable and inexplicable. Thus, if we accept the limitations imposed by Tarski's proposal for avoiding antinomies, we forfeit one of the highest aspirations of the human spirit, the aspiration to self-understanding. (McGee, 1991: 79)

Defense: I fully support McGee's view about the importance of self-understanding and of theories about human thought and action. But I believe that the development of such theories is possible within the Tarskian hierarchy. Indeed, the Tarskian hierarchy is quite hospitable to a large array of theories. It does not forbid us to talk even about truth in a single theory. What it forbids us to do is to evaluate the truth-value of statements of such a theory within that theory. Such a theory can say that human thought is so and so—that it is conducted in a variety of modes, including the immanent, transcendent, and normative modes, that it requires an immanent, transcendent, and normative concept of truth, that this requirement can be spelled out formally using an external hierarchical framework, and so on. What such a theory cannot say is that human thought is so and so *and* it is *true* that human thought is so and so.

Indeed, Tarski's solution to the Liar paradox and the theories he developed within the framework of this solution, such as his theory-definition of truth and his theory-definition of logical consequence, fall under this description. Tarski's theories talk about truth and logical consequence in general. Among other things, they say that truth is hierarchical. What they do not say is that they themselves are true theories of truth and logical consequence (or that so and so follows logically from them). To say these things, we need to transcend these theories (in one way or another) and talk about them from a transcendent standpoint, which, in Tarski's formal rendition, is an external meta-language. Unofficially, a theory can make comments about itself, but not officially.

(e) *Infinite ascent (regress) and absolute generality.*

Criticisms:

Infinite ascent (regress). The truth-statement "S is true", where S belongs to some language L_n, belongs to L_{n+1}. This statement attributes to the L_n-sentence S a property, truth, and to find out (in L_{n+1}) whether S has this property we have to look at what S says and whether its target in the world is as it says it is. Humans, however, are prone to error, and therefore the question whether S is in fact true always arises for them. This question is an L_{n+2} question, and it raises a similar L_{n+3} question, and so on. It seems that we are caught in an infinite ascent, never receiving a final answer to our question.

Absolute generality. To have a full understanding (grasp) of Tarski's definition of truth we have to have a full understanding (grasp) of the Tarskian hierarchy. But we are unable to have such an understanding since we are unable to quantify over the entire Tarskian hierarchy. There are two problems here: (i) There are class-many levels in a Tarskian hierarchy, but Tarskian quantification can encompass only set-many levels. (ii) According to Tarski's theory we cannot say, or learn, anything about truth, without standing, so to speak, within some language in some Tarskian hierarchy. But wherever we stand in a Tarskian hierarchy, we can always ascend to a higher level in that hierarchy, which our present standpoint does not encompass. So it is impossible to grasp the Tarskian hierarchy (hierarchies) in its (their) totality.

Defense:

1. *Infinite ascent.* Although there is no limit on the length of chains of questions about truth, this is not the case with the answers to such questions. A (proper) truth-question, "Is S true?",[28] asked in L_{n+1} about a sentence S of L_n, has a definite (determined, though perhaps unknown) answer that involves at most $n+2$ languages, assuming there is a fact of the matter about that part of the world that S talks about. From this (important) perspective, infinite Tarskian ascent is not problematic.[29]

2. *Absolute generality.* The first problem of absolute generality − inability to quantify over class-many elements − is not special to the Tarskian hierarchy. We are also unable to quantify over the class-many sets there are or the class-many models there are, yet this does not lead us to reject either set theory or model theory. Furthermore, from the point of view of the fundamental principle of truth it is not essential that the Tarskian hierarchy be formulated within a theory of sets, in which a gap between sets and classes arises. (Indeed, historically, Tarski formulated his theory in a Russellian type-theoretic language, in which such a gap does not arise.).[30]

The second problem of absolute generality is the problem of being unable to grasp the entire Tarskian hierarchy (hierarchies), since to do so we have to stand in a place where truth-discourse is formally safe, and this (according to the Tarskian solution) would have to be within some language of some Tarskian hierarchy. In the philosophical literature there are a few proposals for solutions to this or similar problems. Perhaps the most well-known is Quine's (1968) proposal that problems of this kind disappear when we return to our "home language" or "mother tongue". Other solutions, directly related to the Tarskian hierarchy (hierarchies), say that we can grasp the entire Tarskian hierarchy if we do this in a "schematic" (Herzberger, 1970) or "systematically ambiguous" (Parsons, 1974) manner, or if we view the language in which we grasp it as *"sui generis"* (Putnam, 2000).

What solution does the fundamental principle of truth suggest? Clearly not that humans are capable of ascending to a Godly standpoint from which they can see absolutely everything, including the Tarskian hierarchy in its entirety (all the Tarskian hierarchies in their entirety) and in a final way. The fundamental principle talks about *human* ascent or transcendence − ascent or transcendence to a *human*

[28] Where the truth-question "Is S true?" is a *proper* truth-question iff the truth-sentence "S is true" is a proper truth-sentence.

[29] As noted earlier, I disregard here transfinite languages and transfinite hierarchies, which are less significant from the point of view of a philosophical understanding of the concept of truth.

[30] This language has its own weaknesses, but as indicated earlier, it is unlikely that any formal rendering of philosophical ideas like truth, transcendence, knowledge, object, etc. would be *perfect*.

standpoint, *not* to a *Godly* standpoint. The solution the fundamental principle suggests is, accordingly, merely a human solution. And it is a solution *in progress* rather than a final solution. But it is a *substantive* human-solution-in-progress.[31]

This solution to the second absolute-generality problem says that in a significant sense we do grasp the Tarskian hierarchy, and the Tarskian theory of truth formulated within the framework of this hierarchy, in complete generality (complete human generality). One way to broach this solution, or to defend Tarski against criticisms that appeal to the second absolute-generality problem, is to begin with the following observation:

There is no limit to our ability to expand our grasp of the Tarskian hierarchy. No matter in what Tarskian hierarchy, and in what level of this hierarchy, we stand, we can always transcend this standpoint to a higher or broader standpoint from which we view a larger section of the Tarskian hierarchy (hierarchies). This is a well-known observation. But what is new is that this is due to the universality of transcendence. No matter where we stand when we have a certain thought, or make a certain claim, either within a Tarskian hierarchy or elsewhere, we can always transcend this standpoint to a (human) standpoint from which we can reflect upon our thought or claim. This has both a negative and a positive aspect.

Negative aspect: The critic cannot have the final word. The critic, too, is human. The critic, too, speaks from a human standpoint. And given the universality of human transcendence, we can always go beyond the standpoint from which the critic made his most recent criticism to a broader standpoint. No matter how far the critic is chasing the Tarskian, the Tarskian can always extricate herself by a further act of transcendence. (Readers may find some similarity between this claim and Thomas Nagel's (1997) claim that the skeptic can never have the last word.)

Positive aspect: Human transcendence is flexible, both on a formal and—indeed, more so—on an informal level. Human transcendence can take many forms, made possible—but also limited—by human ingenuity. Not only can we transcend any language L_n in a given Tarskian hierarchy to the language L_{n+1} ($=ML_n$), we can also veer sideways to a standpoint from which we understand *the principle* of the Tarskian hierarchy we have left, and in this sense we understand this hierarchy in its totality. This standpoint, for the Tarskian, will be somewhere within another Tarskian hierarchy, but we can transcend this hierarchy too, as well as any other Tarskian hierarchy, and recognize the single principle governing all of them. In this way, we obtain a general understanding of the Tarskian hierarchies without having to quantify over the totality of levels of any, or all, these hierarchies.

In fact, not only can we understand the principle of the Tarskian hierarchies, we can understand the content of the theory of truth that Tarski developed within his hierarchical framework in full generality in a relatively small number of (cognitive) steps. This is due to the fact that the *content* of Tarski's theory does not change from level to level. Its content is essentially *fixed (the same, invariant) across levels (and hierarchies),* and in this sense we fully comprehend it after the first few levels. To see this, let us first briefly reflect on the content of Tarski's theory of truth—the theory Tarski presented as soon as he set down the formal (hierarchical)

[31] For the relevant sense of "substantive", see Sher (2016).

framework for its formulation. Tarski's theory (as explained in Sher, 1999) focuses on the *contribution of logical structure to the truth-value of sentences*, that is, on the way the logical structure of a given sentence affects its truth-value. (This is the reason the Tarskian definition of truth is so valuable in logic, where it serves as a basis for the definition of logical consequence.) The Tarskian definition tells us what the truth-conditions associated with identity, negation, conjunction, …, and existential/universal quantification are. The use of recursion enables the definition to describe, in a finite number of steps, the entire array of truth-conditions associated with the infinitely many logical structures of a given language.

Now, a remarkable thing about Tarski's hierarchical theory of truth is that the truth- (or satisfaction-) conditions for logically-structured sentences (formulas), or the ways different logical structures affect the truth-value of sentences, do not significantly change from one Tarskian language to another. (E.g., the entries for conjunction or universal quantification are essentially the same for all languages.) This is something that anyone who studies Tarski's theory realizes quite quickly, both for different languages on the same level of the hierarchy and for languages on different levels of the hierarchy. (In the latter case, we see this after going through the first few levels of the hierarchy.[32]) So there is an important sense in which by studying Tarski's theory we do study a *single* definition of truth, one that is not affected by the multiplicity of meta-languages.

How to formally express these observations and how far we can go with a formal explanation, is another matter. But it is not something magical or mysterious. We can use means analogous to those used by set theorists to avoid Russell's paradox, which enable them to describe the totality of sets without quantifying over it, or we can find other means.

6 Conclusion

In this paper I have offered a new defense of Tarski's theory of truth and its hierarchical solution to the Liar paradox based on a substantive philosophical principle, "the fundamental principle of truth", and in particular, the immanence and transcendence of truth highlighted by this principle. Tarski's hierarchical solution to the Liar paradox offers a formal rendition of this principle, and while it is possible to improve upon his solution, it is an extremely fruitful, efficient, and, due to its straightforward rendition of the immanence–transcendence duality, philosophically appealing solution.

Acknowledgements I would like to thank the audiences at the *Semantic Paradox and Revenge Workshop,* University of Salzburg, Austria, the National Research University, Higher School of Economics, Moscow, Russia, The *Content, Construction and Convention Workshop,* The Hebrew University of Jerusalem, Israel,

[32] Higher languages in the hierarchy are more powerful than lower languages in ways that do not significantly affect the entries for the logical constants in the definition of truth. (For example, they may be more powerful in having enriched non-logical vocabulary (e.g., mathematical vocabulary), or a higher-level version of the logical vocabulary of the lower languages.)

and the European Network for the Philosophy of Logic, as well as Peter Sher and two anonymous referees, for helpful comments.

References

Beall, J. C., Glanzberg, M., & Ripley, D. (2011/16). Liar paradox. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy*.

Blackburn, S. (1984). *Spreading the word*. Oxford University Press.

Brentano, F. (1995). *Psychology from an empirical standpoint*. O. Kraus (Ed.) (A. C. Rancurello, D. B. Terrell & L. L. McAlister, Trans.), 2nd edn. Routledge.

Herzberger, H. G. (1970). Paradoxes of grounding in semantics. *Journal of Philosophy, 67*, 145–167.

Kirkham, R. L. (1992). *Theories of truth*. MIT.

Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy, 72*, 690–716.

McGee, V. (1991). *Truth, Vagueness and Paradox*. Hackett.

Nagel, T. (1997). *The last word*. Oxford University Press.

Parsons, C. D. (1974). The liar paradox. *Journal of Philosophical Logic, 3*, 381–412.

Putnam, H. (2000). Paradox revisited I: Truth. In G. Sher & R. Tieszen (Eds.), *Between logic and intuition: Essays in honor of Charles Parsons* (pp. 3–15). Cambridge University Press.

Quine, W. V. (1968). Ontological relativity. In W. V. Quine (Ed.), *Ontological relativity and other essays* (pp. 26–68). Columbia University Press.

Quine, W. V. (1970/86). *Philosophy of logic*. Harvard University Press.

Quine, W. V. (1981). Thing and their place in theories. In W. V. Quine (Ed.), *Theories and things* (pp. 1–23). Harvard University Press.

Quine, W. V. (1986). Reply to Harold N. Lee. In L. E. Hahn & P. A. Schillp (Eds.), *The philosophy of W.V. Quine* (pp. 315–318). Open Court.

Quine, W. V. (1995). Reactions. In P. Leonardi & M. Santambrogio (Eds.), *On quine: New essays* (pp. 347–361). Cambridge University Press.

Sher, G. (1999). What Is Tarski's theory of truth? *Topoi, 18*, 149–166.

Sher, G. (2004). In search of a substantive theory of truth. *The Journal of Philosophy, 101*, 5–36.

Sher, G. (2016). *Epistemic friction: An essay on knowledge, truth, and logic*. Oxford University Press.

Siewert, C. (2002/6). Consciousness and intentionality. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy*.

Tarski, A. (1933). The concept of truth in formalized languages. Tarski, 1983, *Logic, semantics, metamathematics.* Hackett, 152–278.

Tarski, A. (1936a). On the concept of logical consequence. Tarski, 1983, *Logic, semantics, metamathematics.* Hackett, 409–420.

Tarski, A. (1936b). The establishment of scientific semantics. Tarski, 1983, *Logic, semantics, metamathematics.* Hackett, 401–408.

Tarski, A. (1983). *Logic, semantics, metamathematics*. Hackett.