Erkenn (2018) 83:899–928
https://doi.org/10.1007/s10670-017-9919-x

# Disquotation and Infinite Conjunctions

**Lavinia Picollo**[1] · **Thomas Schindler**[2]

Received: 31 October 2016 / Accepted: 15 June 2017 / Published online: 10 July 2017


**Abstract** One of the main logical functions of the truth predicate is to enable us to express so-called 'infinite conjunctions'. Several authors claim that the truth predicate can serve this function only if it is fully disquotational (transparent), which leads to triviality in classical logic. As a consequence, many have concluded that classical logic should be rejected. The purpose of this paper is threefold. First, we consider two accounts available in the literature of what it means to express infinite conjunctions with a truth predicate and argue that they fail to support the necessity of transparency for that purpose. Second, we show that, with the aid of some regimentation, many expressive functions of the truth predicate can actually be performed using truth principles that are consistent in classical logic. Finally, we suggest a reconceptualisation of deflationism, according to which the principles that govern the use of the truth predicate in natural language are largely irrelevant for the question of what formal theory of truth we should adopt. Many philosophers think that the paradoxes pose a special problem for deflationists; we will argue, on the contrary, that deflationists are in a much better position to deal with the paradoxes than their opponents.

✉ Lavinia Picollo
Lavinia.Picollo@lrz.uni-muenchen.de

Thomas Schindler
thomas.schindler1980@gmail.com

[1] Munich Center for Mathematical Philosophy, LMU Munich, Germany

[2] Clare College, University of Cambridge, Cambridge, UK

## 1 The Problem

Many philosophers maintain that the truth predicate can serve certain expressive roles of a quasi-logical nature, the most salient of which is to enable us to express so-called 'infinite conjunctions'. These are sentences of the form

$$\text{All } P\text{s are true.} \tag{1}$$

where $P$ is a predicate of the language that applies to infinitely many sentences.[1] They are considered to express the infinitely many $P$s at once or, as is often said, their infinite conjunction, without turning to infinitary or higher-order resources (cf. Quine 1970; Leeds 1978; Putnam 1978; Gupta 1993; Horwich 1998; Field 2007). (The phrase 'expressing infinite conjunctions' is taken from the literature—e.g. Putnam 1978; Gupta 1993—and of course in need of clarification. For the moment, the reader should take it as a technical term.) We call this function of the truth predicate the 'infinite-conjunction' function.

As Quine (1970, chap. 1) points out, the universal quantifier serves a similar purpose. If the infinitely many sentences we want to express differ in one or several individual terms—e.g. "0 is divisible by 2", "2 is divisible by 2", "4 is divisible by 2", etc.—and the class of objects these terms denote is definable in the language by a suitable predicate (e.g. "is an even number"), we can express the infinitely many sentences at once just generalising over those terms using this predicate—e.g. uttering "All even numbers are divisible by 2". However, if the infinitely many sentences we want to express don't differ just in one or more individual terms, this strategy is no longer available. In that case, the truth predicate, interacting with the universal quantifier, might do the job, as long as the sentences at issue share a property definable in the language. For instance, we can assert all theorems of arithmetic at once via "All theorems of arithmetic are true" although they don't even share their logical form, with the aid of the truth predicate. In Quine's (1970, p. 11) own words,

> Where the truth predicate has its utility is in just those places where, though still concerned with reality, we are impelled by certain technical complications to mention sentences. [...] The important places of this kind are places where we are seeking generality, and seeking it along certain oblique planes that we cannot sweep out by generalizing over objects.

Truth theorists also point at other functions of the truth predicate, of a similar nature, such as its epistemic and rhetoric functions. For example, even if there are just *finitely* many sentences we want to express, we can use the truth predicate to *avoid their explicit articulation*—either because we don't know them, or we want to save space or time, or because it is conversationally inappropriate, and so forth. This can be done as above, simply by identifying a predicate they all and only satisfy, or,

[1] Horwich (1998) settles on propositions rather than sentences as truth bearers. For the purposes of this paper, it is irrelevant whether we choose sentences or propositions, since we only deal with eternal sentences in the sense of Quine (1970). If the readers are more inclined towards the latter option, they can understand sentences of the form (1) and the like as ascribing truth to the propositions expressed by the sentences that satisfy predicate $P$, instead of the sentences themselves.

if we're talking about a single sentence, also by choosing a non-quotational name or definite description of it. For example, we can express all four of Maxwell's equations at once uttering

Maxwell's equations are true.

or Gödel's first incompleteness theorem via

Gödel's first incompleteness theorem is true.

without articulating these statements.

Nonetheless, the *logical* interest of truth, with which we are concerned in this paper, lies in its ability to express *infinite* sets of sentences.[2] Unlike the previous examples, when the *P*s are infinitely many (and don't differ just in one or several terms) there's actually *no other way* to express them all at once—at least not in the most commonly used languages, i.e. first-order languages. Thus, as Quine (1970, p. 12) famously said:

> We may affirm the single sentence by just uttering it, unaided by quotation or by the truth predicate; but if we want to affirm some infinite lot of sentences then the truth predicate has its use.

Among the logical functions of the truth predicate, there is also the dual to the infinite-conjunction function, that is, that of expressing infinite disjunctions via sentences of the form "Some *P*s are true". We will return to them later.

It is worth noticing that the distinction between the epistemic and the non-epistemic function of truth *cuts across* the distinction between the logical (more appropriate: infinitary) and the non-logical (finitary) function of truth. One can employ the logical function of truth with or without employing its epistemic function at the same time, and one can employ some non-logical function of truth with or without employing its epistemic function at the same time. For example, when we say that all sentences of the form 'If *p* then *p*' are true, we make use of the infinitary function of truth (arguably) without appealing to its epistemic function, while when we say that all theorems of arithmetic are true, we employ both functions simultaneously (because the set of arithmetical theorems is undecidable). On the other hand, if we express Gödel's first incompleteness theorem using the truth predicate, this might be due to the fact that we cannot remember its exact formulation or because we simply want to save time.

The logical (i.e. infinitary) function of truth encourages the addition of a truth predicate to our logical systems, the formulation of 'logics' of truth, in order to increase our expressive power. More precisely, it prompts the search of formal *theories* of truth, since some syntax theory in the base is needed: truth is in a sense a subject-specific predicate, as it applies only to sentences (or other objects such as propositions or Gödel codes of sentences), and requires their existence as objects (plus some further assumptions about them). Hence, the infinitary function of truth

[2] Moreover, note that singular truth ascriptions can simply be reduced to general truth ascriptions. E.g. "Gödel's first incompleteness theorem is true." can be replaced by "For all $x$, if $x =$ Gödel's first incompleteness theorem, then $x$ is true".

is not of a purely logical nature, but rather logico-linguistic or quasi-logical. We cannot expect our principles governing truth to hold in all models whatsoever but only in those where some syntactic principles hold.

Let $\mathcal{L}$ be a first-order language, the 'base' language, and let $\mathcal{L}_T$ be the result of expanding $\mathcal{L}$ with a monadic predicate symbol $T$ for truth. $\mathcal{L}$ contains a term $\ulcorner\sigma\urcorner$ denoting $\sigma$ (perhaps via some coding) for each expression $\sigma$ of $\mathcal{L}_T$. Let $\Sigma$ be our base theory. We assume $\Sigma$ is able to prove elementary facts about the syntax of $\mathcal{L}_T$, such as e.g. that the concatenation of $\psi$, $\wedge$, and $\chi$ is $\psi \wedge \chi$.[3] A theory of truth is then any (recursively enumerable) subset of $\mathcal{L}_T$ that contains the base theory $\Sigma$ and is closed under first-order (not necessarily classical) logic. A vast number of such theories has been provided so far with the purpose of allowing the truth predicate to fulfil its logico-expressive roles, the infinite-conjunction function being one of them.[4] To properly evaluate what systems are adequate for this latter purpose we would need to answer the following question:

**Question 1** Let $\varphi(x)$ be a predicate that applies only to sentences of $\mathcal{L}_T$. What axioms or rules do we need to postulate for $T$ so that

$$\forall x(\varphi(x) \rightarrow Tx) \tag{InfC}$$

expresses all the $\varphi$s?

A similar question can be raised regarding infinite *disjunctions*. We will briefly say something about them in Sect. 6, but for the moment we will only be concerned with infinite conjunctions, which we consider as more important.

Note that our question does not concern what principles or rules govern the use of the truth predicate in natural language, nor what principles ensure its epistemic and rhetoric functions, but what features of truth account for its infinite-conjunction function in a formal setting.[5] From a deflationist point of view, this is a pressing question: if, as e.g. Horwich (1998, p. 2) argues, "the truth predicate exists *solely* for the sake of a certain *logical* need" (our italics), consisting (among others) in expressing infinite conjunctions, what matters only is the question what truth principles are necessary and sufficient to fulfil this logical need when we devise formal systems or regiment natural language.

The usual answer to this question is that we need a principle governing $T$ that establishes the equivalence between a sentence $\psi$ and its truth predication $T\ulcorner\psi\urcorner$, this is, a *transparency* or *disquotational* principle (cf. Horwich 1998; Halbach 1999; Priest 2006; Field 2008; Beall 2009; Cobreros et al. 2013). The T-schema, given by all the instances of

[3] In the literature on axiomatic theories of truth, one usually takes Peano or Robinson arithmetic as syntax theory, but sometimes also stronger theories such as Zermelo–Fraenkel set theory are considered (e.g. Fujimoto 2012). We assume that $\Sigma$ interprets at least some decent amount of arithmetic.

[4] See Halbach (2011), Field (2008), Beall (2009), and Ripley (2015) for a compendium.

[5] Perhaps not all natural language features are needed. Perhaps we even have to add principles that are not found in natural language.

$$T^{\ulcorner}\psi^{\urcorner} \leftrightarrow \psi \qquad \text{(T-schema)}$$

(called 'T-biconditionals'), where $\psi$ is a sentence of $\mathcal{L}_T$, is probably the most popular, but there are also others, such as the following pair of inference rules:

$$\psi \vdash T^{\ulcorner}\psi^{\urcorner} \qquad \text{(T-Intro)}$$

$$T^{\ulcorner}\psi^{\urcorner} \vdash \psi \qquad \text{(T-Elim)}$$

T-Intro and T-Elim allow us to infer $T^{\ulcorner}\psi^{\urcorner}$ from $\psi$ and vice versa even in hypothetical contexts, unlike the rules NEC and CONEC (cf. Halbach 2011), that allow the inference from $\psi$ to $T^{\ulcorner}\psi^{\urcorner}$ and vice versa only when we have a proof of $\psi$ and $T^{\ulcorner}\psi^{\urcorner}$, respectively.[6]

At this point, of course, a notorious problem arises. Given our syntactic background assumptions, we are able (e.g. via diagonalisation) to formulate liar sentences, sentences that assert their own untruth. These are sentences $\lambda$ for which we can prove

$$\lambda \leftrightarrow \neg T^{\ulcorner}\lambda^{\urcorner}$$

Instantiating the T-schema to $\lambda$, we reach a contradiction in classical logic, known as the 'liar paradox'. This leaves us roughly with two options: we either (1) reject certain instances of our transparency principles, or (2) reject certain classical (meta-)rules of inference. Horwich (1998, pp. 41–42), for example, opts for the first route, remarking that "this restriction need not be severe. It need have no bearing on the propositions of science—the vast majority of which do not themselves involve the concept of truth." However, in his influential book, Field (2008) argues that, without full transparency, the expressive function of the truth predicate is substantially impaired (see especially chapters 7, 8 and 13; we will consider these objections in Sects. 4 and 7 of the present paper).

Nowadays, many philosophers have chosen the second route and adopted some non-classical logic. These logics can be divided into structural and substructural. Roughly, the former can in turn be paracomplete, where the law of excluded middle and the rule of introduction of the conditional fail, or paraconsistent, where inferences such as Ex Falso Quodlibet and Disjunctive Syllogism are invalid.[7] Typically, in paracomplete truth systems the liar sentence is neither true nor false (e.g. Field 2008), while in paraconsistent ones it's regarded as both true and false (e.g. Beall 2009; Priest 2006). On the other hand, substructural approaches impose restrictions on structural properties of the very notion of logical consequence, such as transitivity and contraction (e.g. Cobreros et al. 2013).[8]

[6] In the presence of the deduction theorem, T-Intro and T-Elim taken together imply the T-schema. If Modus Ponens is valid in the theory, the T-schema implies both rules. Thus, while in classical logic the schema and the rules are equivalent to each other, that is not the case in every non-classical system, and there could be reasons to prefer one disquotational principle over the other, as the case may be.

[7] See Priest (2008) for an introduction to structural non-classical logics.

[8] See Restall (2000) for an introduction to substructural logics.

These all seem to be quite drastic moves, so one might reasonably ask what their justification is: Do transparency principles really live up to their promises? In particular, is full transparency necessary (and sufficient) for (InfC) to express all the $\varphi$s? If not, what other rules or principles must $T$ satisfy to let (InfC) express all the $\varphi$s? And what inferences must the background logic validate? (Remember that, at this moment, we are not interested in capturing the notion of truth in natural language, but in the question of what features of truth account for its infinite-conjunction function in a formal setting.) Thus, any answer to Question 1 presupposes (at least partially) an answer to the following:

**Question 2** What does it mean for (InfC) to express all the $\varphi$s?

In the next section, we will look at two answers to Question 2. The first one, which we may call the 'equivalence' account, requires that (InfC) and the $\varphi$s are equivalent to each other (in some suitable sense). The second one, which goes back to Halbach (1999) and may be dubbed the 'finite-axiomatisation' account, requires that (InfC) finitely axiomatises the consequences of the $\varphi$s relative to the truth theory. Both accounts can be seen as intended arguments that transparency principles are indeed needed for (InfC) to express all the $\varphi$s in a formal setting. However, we will argue that none of the them provides a satisfactory answer to Question 2 and moreover, so far as they are correct, they support only the adoption of a principle much weaker than full transparency—a principle that is consistent in classical logic. In Sects. 3 and 4 we show that this principle, together with some regimentation of our truth talk, is sufficient for many (though probably not all) of the quasi-logical functions of the truth predicate. In Sect. 5 we briefly consider the case of infinite disjunctions.

A word of caution. We will not argue that full transparency is not needed for the logical (infinitary) functions of truth. Nor will we argue against the use of non-classical theories of truth. Rather, our aim is to point out that *so far* there are no convincing arguments that transparency is needed for the logical functions of truth, and that (with a little bit of regimentation) many of the functions of truth can in fact be performed using principles much weaker than transparency. This can be seen (a) as a plea for a more thorough investigation of what the several functions of truth consist of and what principles each of them requires, and (b) as a challenge for proponents of non-classical theories to provide arguments for transparency.

In Sect. 6 we argue that, from a deflationist point of view, the principles that govern the use of the truth predicate in natural language are to a certain extent irrelevant for the function of truth in a (semi-)formal setting and, moreover, that the paradoxes do not pose a special problem for deflationism. In Sect. 7 we consider a possible problem to our approach. In Sect. 8 we conclude by summarising our findings. In the Appendix, we prove some of the more technical claims made in this paper.

## 2 The Equivalence and the Finite-Axiomatisation Accounts

Clearly, transparency guarantees that, for instance,

Gödel's first incompleteness theorem is true.

expresses Gödel's first incompleteness result, for it establishes an *equivalence* between these two sentences (given the additional premise that 'Gödel's first incompleteness result' refers to Gödel's first incompleteness result). Analogously, transparency suffices to conclude that

Maxwell's equations are true.

is equivalent to the conjunction of Maxwell's equations, and therefore expresses them (given the additional premise that 'Maxwell's equations' applies to Maxwell's equations). These examples seem to suggest that (1) some transparency principle is needed to secure the equivalence between a set of (possibly infinitely many) sentences we want to express and the truth ascription we use for this purpose, and (2) this equivalence is precisely what it means for a truth ascription to express some given set of sentences. This is what we call the 'equivalence' account. Some version of this account seems to be proposed in e.g. Putnam (1978, p. 15), Gupta (1993, pp. 60–61), Horwich (1998, p. 3).

The question of course is what 'equivalence' amounts to here. Perhaps the most natural interpretation of 'equivalence' is *mutual implication* in some suitable sense. That is, the sentence

$$\forall x(\varphi(x) \rightarrow Tx) \tag{InfC}$$

implies and is implied (in some suitable sense) by the set of all objects falling under the predicate $\varphi(x)$ or, better, the set of all its 'instances', that is the set of all sentences of the form

$$\varphi(\ulcorner\psi\urcorner) \rightarrow \psi,$$

which has the advantage that we don't need to know what the $\varphi$s are.

It is immediately clear then that 'mutual implication' cannot be taken in the usual first-order sense. As is well known, universally quantified claims are stronger than the collection of their instances. While it is in principle possible to infer all sentences of the form $\varphi(\ulcorner\psi\urcorner) \rightarrow \psi$ from (InfC) given a transparency principle, the other direction of the implication is blocked by compactness, unless some finite subset of $\{\varphi(\ulcorner\psi\urcorner) \rightarrow \psi : \psi \in \mathcal{L}_T\}$ already implies (InfC).[9]

This problem is of course reminiscent of an old, well-known problem: the T-biconditionals are too weak to prove any non-trivial generalisations about truth, as Tarski (1935) has already noted. In particular, as Halbach (1999, proposition 1) has shown (for the typed case), whenever there is a model in which $\varphi(x)$ applies to infinitely many objects, then $\forall x(\varphi(x) \rightarrow Tx)$ cannot be derived using T-biconditionals as the only truth principles, not even in the presence of induction principles.

[9] There are further intuitive differences between expressions like (InfC) and real infinite conjunctions. In some cases, the former have more expressive power than the latter. For example, turning to (InfC) we can express all instances of any schema in the language at once, i.e. the $\varphi$s, including the instance given by (InfC) itself; whilst the well-foundedness of a well-formed formula in infinitary languages precludes this, as formulae are not allowed to contain themselves as subformulae.

Could adopting some $\omega$-rule-like rule be a solution?[10] If the $\omega$-rule held, together with a principle of disquotation, the set of all sentences of the form $\varphi(\ulcorner\psi\urcorner) \rightarrow \psi$ would imply (InfC). Unfortunately, we can never actually apply the $\omega$-rule (not even constructive versions, for that matter).[11] We would have to check infinitely many premisses before drawing a conclusion. Therefore, while in theory the $\omega$-rule appears to lead to the desired equivalence, it is absolutely useless in practice; it makes no difference in the inferences we can actually draw.[12] One may object that while the $\omega$-rule is indeed useless from a practical point of view, it does not follow that one cannot use it to explicate what it means for a generalisation to express an infinite set of sentences. But in that case the equivalence in question would no longer be 'fixed' by our use of language.

Moreover, if we *could* apply the $\omega$-rule, the truth predicate would lose a lot of its interest, which lies in its ability to finitely express infinite sets of sentences. If the possibility of reasoning with infinitely many premisses were already given, the need for such a device would diminish considerably. Its interest, as we mentioned at the beginning of this paper, rests on the ubiquity of finitary logics, both in science and in philosophy. While there still could be reasons for having a truth predicate in some infinitary logic, these would not be the same reasons that originally prompted our search for a formal theory of truth. Thus, in the remainder of this paper we will restrict our attention to finitary logics.

It is of no help either to try to cash out the equivalence in terms of some intended model. Such a proposal would be something along the following lines: instead of requiring that the truth principles we adopt guarantee the mutual implication between (InfC) and $\{\varphi(\ulcorner\psi\urcorner) \rightarrow \psi : \psi \in \mathcal{L}_T\}$ in every model of the base theory, all that is needed is that this mutual implication holds in all extensions of the *standard* model of the base language $\mathcal{L}$ to the whole $\mathcal{L}_T$ satisfying our truth principles. For simplicity, let us focus on the typed case, where disquotation is restricted to sentences of $\mathcal{L}$ (without the truth predicate). Let $\mathcal{M}$ be $\mathcal{L}$'s intended model. In $\mathcal{M}$, the interpretation of $\ulcorner\psi\urcorner$ is (the code of) $\psi$ for every sentence $\psi$ in $\mathcal{L}$. Let $\varphi(x)$ be a predicate of $\mathcal{L}$ whose extension in $\mathcal{M}$ is the set $\Phi \subseteq \mathcal{L}$, and let $(\mathcal{M}, S)$ extend $\mathcal{M}$ to $\mathcal{L}_T$, assigning $S$ to $T$ as its extension. Then, according to this proposal, a truth ascription of the form (InfC) expresses the $\varphi$s if and only if for every model $(\mathcal{M}, S)$ of $\mathcal{L}_T$ satisfying the 'right' truth principles, the following equivalence holds:

$$(\mathcal{M}, S) \models \forall x(\varphi(x) \rightarrow Tx) \quad \text{if and only if,} \quad \text{for all} \quad \psi \in \Phi, \quad \mathcal{M} \models \psi.$$

The problem with this proposal is that it can only guarantee the aforementioned equivalence between (InfC) and the set of $\varphi$s under the *intended* interpretation of the non-logical vocabulary of $\mathcal{L}$. But this is obviously just the semantic version of the previous proposal involving the $\omega$-rule, which we just rejected. The set of all

[10] $\omega$-rule-like rules allow us to infer universally quantified statements from the set of all their instances. Recall that such rules are semantically valid e.g. in second-order arithmetic (with standard semantics).

[11] Constructive versions of the $\omega$-rule impose conditions on the infinite set of premises to which the rule is applied, e.g. that the set of premises is recursive (computable) or that there is a 'uniform way' of proving them, as in Baker et al. (1992).

[12] Raatikainen (2005) makes a similar point in a slightly different context.

sentences true in some intended model will be highly complicated. We are after rules for reasoning, and truth in the intended model cannot be captured by them.

Moreover, given the quasi-logical nature of the function of the truth predicate we want to capture, the original aim of this paper was to formulate 'logics' of truth, theories that apply across the board—the truth predicate ought to fulfil its function regardless of the interpretation of the base language we have in mind (as long as these interpretations satisfy the axioms of the syntax theory). In order to see why this constraint is important, suppose for example that our base language $\mathcal{L}$ is the language of set theory. It is doubtful, to say the least, that we have a clear grasp (concept) of the intended model of set theory. Even more, current trends in the philosophy of mathematics suggests that there might not be such a thing as 'the' intended model of set theory (e.g. the multiverse view of sets).

Let us therefore turn to an alternative account of what it means for (InfC) to express all the $\varphi$s, which is due to Halbach (1999). In fact, it is the only formally worked out answer to Question 2 we are aware of, and it seems to avoid the problem posed by the compactness of first-order logic. According to this proposal, (InfC) and the set of conditionals of the form $\varphi(\ulcorner\psi\urcorner) \rightarrow \psi$ have to be equivalent only with respect to their consequences *in the language of the* $\varphi$*s*. In effect, this condition requires that (InfC) *finitely axiomatises* the set of conditionals *relative to the truth theory*. Thus we refer to Halbach's account as the 'finite-axiomatisation' account. That (InfC) finitely axiomatises the conditionals $\varphi(\ulcorner\psi\urcorner) \rightarrow \psi$ is *prima facie* a plausible explication of what it means for (InfC) to express them.

To avoid paradoxes and simplify the matter, Halbach restricts his attention to sentences that do not contain the truth predicate. He shows that the T-biconditionals for these sentences are *sufficient* to guarantee that (InfC) finitely axiomatises the set of the conditionals $\varphi(\ulcorner\psi\urcorner) \rightarrow \psi$, in the following sense. Recall that $\mathcal{L}$ is the $T$-free fragment of $\mathcal{L}_T$.

**Proposition 3** (Halbach) *Let* $\Gamma \subseteq \mathcal{L}_T$ *extend a syntax theory* $\Sigma \subseteq \mathcal{L}$ *with the* T-schema*, (or* T-Intro *and* T-Elim*) restricted to* $\mathcal{L}$*, and let* $\varphi(x)$ *be a predicate of sentences of* $\mathcal{L}$*. Then* $\Sigma + \{\varphi(\ulcorner\psi\urcorner) \rightarrow \psi : \psi \in \mathcal{L}\}$ *and* $\Gamma + \forall x(\varphi(x) \rightarrow Tx)$ *have exactly the same T-free consequences (in classical logic).*[13]

Halbach (2011, pp. 59–60) comments on this result as follows:

> [This proposition] shows that infinite generalizations understood as schemata of the form $\varphi(\ulcorner\psi\urcorner) \rightarrow \psi$ can be expressed by a single sentence in the presence of the disquotation sentences. In a sense the infinitely many sentences $\varphi(\ulcorner\psi\urcorner) \rightarrow \psi$ have been replaced by the single sentence $\forall x(\varphi(x) \rightarrow Tx)$ and the infinitely many disquotation sentences.

The fact that we have replaced one infinite set of sentences (the sentences of the form $\varphi(\ulcorner\psi\urcorner) \rightarrow \psi$) with another (the T-biconditionals) does not undermine Halbach's proposal. In his own words:

[13] Cf. Halbach (1999, proposition 2). For further discussion on this result see Heck (2004) and Kemp (2005). Note: when we say "let $\varphi(x)$ be a predicate of sentences of $\mathcal{L}$", we mean that the syntax theory $\Sigma$ proves (the formalisation of) the claim "For all $x$, if $\varphi(x)$ then $x$ is a sentence of $\mathcal{L}$".

> I think it would be coherent to claim that the disquotation sentences are in the 'background' in very much the same way as rules of inferences (such as modus ponens) are in the background, as logic cannot be axiomatized without axiom schemata or rules with infinitely many instances. (Halbach 2011, pp. 60–61)

Interestingly, Halbach's account does not justify the adoption of transparency principles but only of a much weaker principle, namely T-Elim or, alternatively, the left-to-right direction of the T-schema, this is,

$$T\ulcorner\psi\urcorner \rightarrow \psi \tag{T-Out}$$

For the following proposition shows that Halbach's result does not rely on the 'introduction' half of transparency principles at all. In other words,

$$\psi \rightarrow T\ulcorner\psi\urcorner \tag{T - In}$$

or T-Intro, does not play any role in the proof of Proposition 3.

**Proposition 4** *Let $\Gamma \subseteq \mathcal{L}_T$ extend a syntax theory $\Sigma \subseteq \mathcal{L}$ with T-Out (or T-Elim) restricted to $\mathcal{L}$, and let $\varphi(x)$ be a predicate of sentences of $\mathcal{L}$. Then $\Sigma + \{\varphi(\ulcorner\psi\urcorner) \rightarrow \psi : \psi \in \mathcal{L}\}$ and $\Gamma + \forall x(\varphi(x) \rightarrow Tx)$ have the same T-free consequences (in classical logic).*

*Proof* Obviously, if $\chi \in \mathcal{L}$ is a consequence of $\Gamma + \forall x(\varphi(x) \rightarrow Tx)$, then $\chi$ is also a consequence of $\Gamma$+T-In+$\forall x(\varphi(x) \rightarrow Tx)$, the result of extending $\Sigma$ with the full T-schema for sentences of $\mathcal{L}$. Thus, by Halbach's result, $\chi$ is also a consequence of $\Sigma + \{\varphi(\ulcorner\psi\urcorner) \rightarrow \psi : \psi \in \mathcal{L}\}$. Conversely, let $\chi \in \mathcal{L}$ be a consequence of $\Sigma + \{\varphi(\ulcorner\psi\urcorner) \rightarrow \psi : \psi \in \mathcal{L}\}$. By compactness, only finitely many sentences in $\{\varphi(\ulcorner\psi\urcorner) \rightarrow \psi : \psi \in \mathcal{L}\}$ have been used in its derivation. Clearly, all of them follow from $\forall x(\varphi(x) \rightarrow Tx)$ plus the relevant instances of T-Out. □

Thus, *if* the finite-axiomatisation account gives the right answer to Question 2, it does not support the adoption of introduction principles for truth, but only for elimination principles.

Halbach's proposal is formulated for the typed case only, but the truth predicate, as a device for infinite conjunctions, is there to express any infinite set of sentences definable in the language, not just the ones that only contain $T$-free sentences. For example, if we assert that all sentences of the form $\psi \rightarrow \psi$ are true, then we want to include *all* sentences of our language in the range of our quantifiers, not just those from the base language. Unfortunately, it is not clear how to extend the finite-axiomatisation criterion to the type-free case. If (InfC) and the set of its 'instances' were to have the same consequences in the language with the truth predicate, then the latter would have to imply the former, since (InfC) is a consequence of itself. But, as before, in finitary languages in most cases the set of consequences of $\{\varphi(\ulcorner\psi\urcorner) \rightarrow \psi : \psi \in \mathcal{L}_T\}$ is just a proper subset of (InfC)'s.

*Remark* Proposition 3 shows how to axiomatise the infinite set $\{\varphi(\ulcorner\psi\urcorner) \rightarrow \psi : \psi \in \mathcal{L}\}$ by the sentence $\forall x(\varphi(x) \rightarrow Tx)$, using the T-biconditionals for sentences of

$\mathcal{L}$ in the background. In their well-known paper, Craig and Vaught (1958) have shown something stronger. Assume that $\Gamma$ is a theory in the language $\mathcal{L}$ that has only infinite models. Then, if $\Gamma$ is axiomatisable, then $\Gamma$ is finitely axiomatisable using additional predicates. Suppose, for example, that $\Gamma$ is Peano arithmetic. Let $\Gamma^*$ be the theory consisting of (a) some finitely axiomatised syntax theory,[14] (b) the compositional truth axioms for the language of PA (which are also finitely many), and (c) the sentence that all theorems of Peano arithmetic are true. Then $\Gamma^*$ is a finite axiomatisation of PA. The compositional truth axioms are a formalisation of Tarski's inductive truth definition, such as

$$\forall x \forall y (Tx \dot{\wedge} y \leftrightarrow Tx \wedge Ty)$$

or

$$\forall x (T \dot{\neg} x \leftrightarrow \neg Tx)$$

where $\dot{\wedge}$ is a function symbol for the function that maps (the code of) two formulae to (the code of) their conjunction, and $\dot{\neg}$ is a function symbol for the function that maps (the code of) a formula to the (code of) its negation. Now, suppose that $\psi$ is a theorem of PA. Then $\Gamma^*$ knows this by (a). By (c), $\Gamma^*$ proves $T\ulcorner\psi\urcorner$. Now, using induction in the metalanguage, one derives the T-biconditionals from the compositional axioms. Hence, one can infer $\psi$ from $T\ulcorner\psi\urcorner$. Essentially, the argument requires a finite axiomatisation of the infinitely many T-biconditionals—this is the job of the compositional axioms. However, it is clear that the argument does not require the full T-biconditionals, but only their elimination half. Since the compositional clauses were only needed for a finite axiomatisation of the T-biconditionals, it is clear that the argument does not rely on them either. The Craig–Vaught result does not require the full compositional truth axioms, but only their left-to-right direction—we only need to finitely axiomatise T-Out.

## 3 The Elimination Property

We do not want to suggest that the finite-axiomatisation account gives the right answer to Question 2, but it is clear that it captures something important about the infinite-conjunction function. We have remarked that the finite-axiomatisation account is not easily extended to the type-free case. However, while in general

$$\forall x (\varphi(x) \rightarrow Tx) \qquad \text{(InfC)}$$

and the set of conditionals of the form $\varphi(\ulcorner\psi\urcorner) \rightarrow \psi$ cannot have *exactly* the same consequences, we may expect that all consequences that can be deduced from the conditionals are deducible from (InfC). If that were the case, (InfC) could still be

[14] If the language of the syntax theory overlaps with the language of the theory to be finitely axiomatised, we need to formulate the syntax theory in a copy of the language, so that $\Gamma^*$ remains conservative over $\Gamma$. For a definition of conservativity, see footnote 15.

said to finitely axiomatise the conditionals in a broader sense. In Horwich's (1998, p. 124) own words, "[the infinite-conjunction] function of truth requires merely that the generalizations permit us to *derive* the statements to be generalized".

We say of a truth predicate satisfying these requirements that it has the *elimination property*. Formally, the inference from $\forall x(\varphi(x) \rightarrow Tx)$ and $\varphi(\ulcorner\psi\urcorner)$ to $\psi$ should hold, this is:

$$\forall x(\varphi(x) \rightarrow Tx),\ \varphi(\ulcorner\psi\urcorner) \vdash \psi$$

If the deduction theorem holds for $\vdash$, the elimination property ensures that

$$\forall x(\varphi(x) \rightarrow Tx) \vdash \varphi(\ulcorner\psi\urcorner) \rightarrow \psi \tag{2}$$

Thus, if $\vdash$ is transitive, everything deducible from the infinitely many sentences $\varphi(\ulcorner\psi\urcorner) \rightarrow \psi$ is deducible from $\forall x(\varphi(x) \rightarrow Tx)$.

As a consequence, the following holds in classical logic: if the truth theory has the elimination property and is also *conservative* over its base theory,[15] then the generalisation will have exactly the same $T$-free consequences as the set $\{\varphi(\ulcorner\psi\urcorner) \rightarrow \psi : \psi \in \mathcal{L}\}$. Thus, given conservativity, the elimination property implies finite axiomatisability in the sense of Halbach as a special case.

While one may doubt Horwich's thesis that the infinite conjunction function "merely" requires the elimination property, it is clear that the elimination property is highly desirable for a logic of truth. If you have committed yourself to "All theorems of arithmetic are true" and "$2 + 2 = 4$" is a theorem of arithmetic, then you have committed yourself to "$2 + 2 = 4$".

As the reader will anticipate, in order to ensure the elimination property only one half of a disquotational principle is needed, namely, the 'elimination' half.

**Observation 5** Let $\Gamma \subseteq \mathcal{L}_T$ be a truth theory where T-Out or T-Elim hold, formulated over a logic in which the rules of Universal Instantiation and *Modus Ponens* are valid. Then $T$ has the elimination property in $\Gamma$.

The proof of this observation is trivial. A consequence of it is that in such logics T-Out or, equivalently, T-Elim, is *sufficient* for the elimination property. This includes classical logic, where both T-Out and T-Elim are not only consistent but also conservative over the usual syntax theories. Thus, *to have a truth predicate with the elimination property there is no need to weaken classical logic.*

Under fairly general circumstances, T-Elim and T-Out are not only sufficient but also *necessary* conditions for granting $T$ the elimination property.

**Observation 6** Let $\Gamma \subseteq \mathcal{L}_T$ be a truth theory formulated over a logic where identity behaves classically,[16] and the usual rules of introduction of the conditional and the universal quantifier are valid. If $\Gamma$ has the elimination property, then T-Elim and T-Out hold in $\Gamma$.

[15] A theory of truth $\Gamma$ is conservative over its base theory $\Sigma$ if and only if for every $T$-free sentence $\psi$, if $\Gamma$ proves $\psi$ then so does $\Sigma$.

[16] This is, the inference from $s = t$ and $\varphi(s)$ to $\varphi(t)$ holds for every formula $\varphi(x)$ and pair of terms $s$, $t$.

*Proof* Assume $T\ulcorner\psi\urcorner$ and that $x = \ulcorner\psi\urcorner$. By the laws of identity, we get that $Tx$. By the introduction rules for the conditional and the universal quantifier, we know that $\forall x(x = \ulcorner\psi\urcorner \rightarrow Tx)$. Thus, the elimination property together with the fact that $\ulcorner\psi\urcorner = \ulcorner\psi\urcorner$ gives us $\psi$. A further application of the introduction of the conditional delivers T-Out. □

It is worth noticing that certain non-classical truth theories in which transparency principles hold unrestrictedly do not have the elimination property. To start with a simple example, consider the theory that results from adding a transparency principle to the logic of paradox LP (cf. Priest 2006) formulated in $\mathcal{L}_T$ plus some syntax theory. As is well known, LP doesn't satisfy Modus Ponens, so observation 5 does not apply. One can show there is no way of guaranteeing the inference from (InfC) to each conditional of the form $\varphi(\ulcorner\psi\urcorner) \rightarrow \psi$ .

**Observation 7** Let $\Gamma \subseteq \mathcal{L}_T$ be a truth theory formulated over LP that extends the syntax theory $\Sigma$ with the unrestricted T-schema and rules T-Intro and T-Elim. Then, $\Gamma$ does not have the elimination property.[17]

Of course, this should not come as a surprise. With the notable exception of Beall (2013), the failure of such a basic and intuitive rule as Modus Ponens in LP led many paraconsistent-minded philosophers (e.g. Priest 2006, chap. 6; Beall 2009) to search for a 'suitable conditional', this is, a conditional-like connective that could be added to LP, satisfying not only Modus Ponens but also other prima facie desirable principles. While many of these new theories grant the truth predicate the elimination property, some of them fail to satisfy the contrapositive of it, which seems equally desirable. For instance, Priest (2006) adopts a non-contraposable conditional with which he formulates the T-schema. Since the new conditional satisfies Modus Ponens, Modus Tollens no longer holds, and therefore the inference

$$\varphi(\ulcorner\psi\urcorner), \neg\psi \vdash \neg\forall x(\varphi(x) \rightarrow Tx)$$

isn't generally valid. Similarly, Kripke's fixed-point theory (Kripke 1975) with the Weak Kleene scheme fails to satisfy this rule of inference.[18]

These observations casts severe doubt on the adequacy of the theories at issue as systems intended to provide a device for expressing infinite conjunctions. Presumably, the original point of weakening classical logic was to accommodate a disquotational principle avoiding triviality, to allow the truth predicate to express infinite conjunctions—perhaps among other things. But in the cases considered, the weakening is so severe that the logic can no longer guarantee the elimination property. Of course, this is not an argument against the use of non-classical truth theories per se. There are many non-classical truth theories that satisfy both the elimination property and its contrapositive. What we would like to highlight is that transparency does not ensure the expression of infinite conjunctions by itself.

There seems to be an obvious reason why a truth predicate governed by T-Out alone is not enough. Namely, there are predicates satisfying T-Out, such as $x \neq x$,

[17] A proof can be found in the Appendix.

[18] A proof of these observations can be found in the Appendix as well.

that are definable in every (classical) first-order theory. However, there is an important difference between (a) being able to define a predicate satisfying T-Out and (b) adding a primitive predicate governed by T-Out. While a primitive truth predicate governed by T-Out as its sole axiom scheme *can* be interpreted by the empty set, it is consistent to assume that its extension is non-empty. This is not possible for the predicate $x \neq x$. Therefore, a primitive truth predicate governed by T-Out enables us to consistently assume, in the course of a hypothetical argument, that e.g. "For all $x$(if $x$ is a sentence of the form '$P \vee \neg P$' then $x$ is true". Obviously, this is not possible if '$x$ is true' is defined as $x \neq x$. In this sense, a truth predicate governed by T-Out is not definable in or reducible to the base theory.

Nevertheless, we would not want to suggest that T-Out, or T-Elim, *alone* qualifies as a suitable truth theory. Firstly, we do not believe that the finite-axiomatisation account gives the correct answer to Question 2. Secondly, if we want to be able to prove certain generalisations, such as "For all $x$, if $x$ is a sentence of the form '$P \rightarrow P$', then $x$ is true", then this will require principles beyond T-Out. Thirdly, in the next section we will present other possible reasons to adopt additional truth principles.

## 4 What About Introduction Principles?

We have argued that the elimination property is an important aspect of the infinite-conjunction function. (We have refrained, however, from identifying the two.) One might now wonder whether introduction principles, such as T-In or T-Intro, contribute anything to the utility of truth. Does the infinite-conjunction function presuppose them? As long as we don't have a sound and complete account of what expressing infinite conjunctions means, this question cannot be properly answered. What is clear, however, is that the accounts considered so far don't support the need of adopting such principles. Therefore, it might be instructive to consider arguments of a different kind in favour of the adoption of introduction principles.

We have repeatedly noted that introduction principles for truth do not suffice to introduce sentences of the form

$$\forall x(\varphi(x) \rightarrow Tx) \tag{InfC}$$

if the $\varphi$s are infinitely many. However, suppose there are only finitely many sentences falling under $\varphi(x)$, say $\psi_1, \ldots, \psi_n$. What the introduction principles for truth would allow us to do is to derive the generalisation (InfC) from $\psi_1, \ldots, \psi_n$ given the premise that $\varphi(x)$ applies exactly to $\psi_1, \ldots, \psi_n$.[19] Does this give us reasons to adopt introduction principles? We don't think so. Note that in the envisaged case, the introduction of the generalisation (InfC) is logically dispensable. If we wish to assert finitely many sentences $\psi_1, \ldots, \psi_n$, we can simply assert their conjunction. So this does not give us any *logical* reasons to adopt introduction principles for truth—it does not concern the logical (i.e. infinitary) function of truth that we are interested in.

[19] More precisely, the assumption is $\forall x(\varphi(x) \leftrightarrow x = \ulcorner\psi_1\urcorner \vee \cdots \vee x = \ulcorner\psi_n\urcorner)$.

Let us therefore have a look at some other arguments for introduction principles that can be found in the literature. Consider the following scenario by Field (2008, p. 210) where a sentence of the form (InfC) occurs in the *antecedent* of a conditional. Suppose you do not remember exactly what Jones said, but you believe that it entails a certain proposition $\psi$. Thus, in order to express your belief you might say "If everything that Jones said is true then $\psi$", that is,

$$\forall x(\chi(x) \rightarrow Tx) \rightarrow \psi \tag{3}$$

where $\chi(x)$ applies exactly to the sentences uttered by Jones. Then, relative to the assumption that what Jones said is exactly $\psi_1, \ldots, \psi_n$, we want the above to imply

$$\psi_1 \wedge \cdots \wedge \psi_n \rightarrow \psi \tag{4}$$

As Field notices, to derive (4) from (3) and the information that $\psi_1, \ldots, \psi_n$ are the only $\chi$s, $\psi_i$ must entail $T\ulcorner\psi_i\urcorner$, this is, an introduction rule is needed. Since elimination rules are also required, Field concludes, classical logic must be abandoned.

Again, in this case the use of the truth predicate has no particular *logical* interest. It isn't logically indispensable, since the sentences we want to express are finitely many, namely, $\psi_1, \ldots, \psi_n$. Thus, we believe it does not give us enough reasons for adopting introduction principles. What is more, there is an alternative and easy way to deal with the Jones case that only involves an elimination principle. For we can express (4) with a simple generalisation of the form (InfC), instead of (3). Let $\varphi(x)$ be the predicate '$x$ is the unique sentence obtained by concatenating the conjunction of the $\chi$s with '$\rightarrow \psi$"'. Then, we can choose (InfC) to express (4): by Observation 5, in any classical T-Out or T-Elim theory (that contains enough syntax theory to prove basic facts about concatenation) we can derive the latter from the former, relative to the assumption that what Jones said is exactly $\psi_1, \ldots, \psi_n$.

The strategy works for infinite cases as well. Consider the claim that every truth is knowable, formally:

$$\forall x(Tx \rightarrow \Diamond Kx) \tag{5}$$

where $Kx$ means that $x$ is known and $\Diamond$ is the possibility operator. Certainly, we want the above to imply all sentences of the form

$$\psi \rightarrow \Diamond K\ulcorner\psi\urcorner \tag{6}$$

Again, (5) won't give us all instances of (6) unless we have some introduction principle at our disposal. But we can simply apply the same trick as before and use another generalisation in order to capture all instances of (6), namely, "All instances of (6) are true". Given an elimination principle (plus some syntax theory), the latter will yield all instances of (6), as desired.

More generally, assume we want to capture all instances of a schema '$\ldots\psi\ldots$' for each sentence $\psi$ in the language. Instead of replacing $\psi$ with $Tx$ and then quantifying over $x$—i.e. instead of $\forall x \ldots Tx \ldots$—we could say that every instance of '$\ldots\psi\ldots$' (where $\psi$ is a sentence) is true, using a formula of the form (InfC). By an

elimination principle such as T-Out or T-Elim (plus some syntax theory), this will imply all the statements that we initially wanted to capture.

Again, note that what is at issue here is not to account for e.g. the intuitive validity of the inference from (5) to (6) (which may hold in natural language), but to provide a device that allows us to finitely capture the infinitely many instances of (6) in a (semi-)formal setting. *This* is what we are interested in in this paper.

We concede that there are some limitations to the above method, and this suggests that principles beyond T-Out might be needed for a satisfactory formal theory of truth. For example, if we combine (5) with the claim that all theorems of Peano arithmetic are true, we can deduce that all theorems of Peano arithmetic are knowable, i.e.

$$\forall x(Bew(x) \rightarrow \diamond Kx) \tag{7}$$

where $Bew(x)$ is a provability predicate for Peano arithmetic. On the other hand, if we combine "All instances of (6) are true" with the claim that all theorems of arithmetic are true, we can only deduce the instances of (7), but not (7) itself.

Let us have a look at another example. Consider for a moment the standard definition of knowledge. An agent is said to know a sentence just in case she believes it, she is justified in doing so, and, moreover, the sentence is true (and some Gettier condition is satisfied). Formally, epistemologists assert[20]

$$\forall x(K(x) \leftrightarrow C(x) \wedge Tx) \tag{8}$$

which is intended to capture the infinitely many instances of the schema

$$K(\ulcorner\psi\urcorner) \leftrightarrow C(\ulcorner\psi\urcorner) \wedge \psi \tag{9}$$

where $C(x)$ resumes all conditions for knowledge except truth. As before, we can capture all instances of (9) by saying "All instances of (9) are true", instead of using (8). As a definition, however, this has its shortcomings. The predicate $K$ is no longer eliminable, and the definition does not satisfy the condition of being non-creative.

Nonetheless, we believe one should carefully weight the costs of this against introducing a non-classical truth predicate into the definition of knowledge. The non-classicality of truth can be 'contagious': it might spread out and turn knowledge into a non-classical predicate too. For example, if the law of excluded middle is rejected for some sentences that contain semantic vocabulary then this will affect the knowledge predicate too. In the envisaged case, presumably it won't follow that we either know or don't know that the liar is true. Prima facie, an analogous point can be made for any predicate of sentences (or propositions) involving the truth predicate in its definition.

We have seen that there are good reasons for adopting elimination principles for truth, whereas we found that the arguments for adopting introduction principles are not entirely convincing—at least when we only consider the infinite-conjunctions function. While our findings may not be conclusive, they present a challenge for all

[20] A discussion of the Gettier problem with knowledge as a predicate is found in Halbach (2016) (with references to further literature). Huemer (2005) is also useful.

those that believe that introduction principles for truth must be part of a logic of truth.

## 5 Infinite Disjunctions

Let us now briefly consider sentences of the form "Some *P*s are true" or, more formally,

$$\exists x(\varphi(x) \wedge Tx) \qquad \text{(InfD)}$$

which are often said to express *infinite disjunctions*. We have seen that elimination principles are highly desirable for a reasonable logic for expressions of the form "All *P*s are true". Since "Some *P*s are true" seems to be dual to "All *P*s are true", one might think that the former call for introduction principles for truth.

A natural thought is that, at the very least, our principles for truth should allow us to infer (InfD) from $\varphi(\ulcorner\psi\urcorner) \wedge \psi$. And this would involve an introduction principle for truth. However, it is by no means obvious that this inference is indispensable for the logical function of truth. The reason is that the conclusion is simply a weakening of the the first sentence. Our interest in the truth predicate does not consist primarily in the sentences that contain the truth predicate, but in the sets of sentences that we cannot express without the help of the truth predicate. In the case under consideration, we already start with the sentence without the truth predicate. What would be the point of introducing the latter?

However, let us have a look at what expressing an infinite disjunction could mean. One way to explicate this would be via intended models. According to this proposal, (InfD) expresses the infinite disjunction of the $\varphi$s if and only if the following equivalence holds:

$$(\mathcal{M}, S) \models \exists x(\varphi(x) \wedge Tx) \text{ if and only if, for some } \psi \in \Phi, \quad \mathcal{M} \models \psi$$

where $\mathcal{M}$ is the intended model of the base language, $\Phi$ is the extension of $\varphi$ in $\mathcal{M}$, and $S$ is the extension assigned to the truth predicate. Since we have already rejected the corresponding account for infinite conjunctions, we need not go into any details here.

Let us therefore have a look at how Halbach's finite-axiomatisation account deals with infinite disjunctions. Recall (a) that a sentence involving the truth predicate was said to express the infinite conjunction of some sentences if and only if they have the same *T*-free consequences. In order to deal with infinite disjunctions, Halbach makes the following additional assumptions: (b) that the infinite disjunction of the $\varphi$s is equivalent to the infinite conjunction of the negation of the $\varphi$s (in the sense that they both have the same consequences in the language $\mathcal{L}$); (c) that a sentence $\chi$ involving the truth predicate expresses an infinite disjunction if and only if $\neg\chi$ expresses the negation of the infinite disjunction.

Now consider the infinite disjunction $\bigvee\{\varphi(\ulcorner\psi\urcorner) \wedge \psi : \psi \in \mathcal{L}\}$. By assumption (b), its negation is equivalent to the infinite conjunction $\bigwedge\{\varphi(\ulcorner\psi\urcorner) \rightarrow \neg\psi : \psi \in \mathcal{L}\}$. Halbach shows that given a *transparency* principle for truth, the latter has the same

$T$-free consequences as the sentence $\forall x(\varphi(x) \rightarrow \neg Tx)$, and is therefore expressed by the latter, because of (a). Thus, by (c), the negation of that sentence expresses the infinite disjunction of the $\varphi$s.

The argument just given relies essentially on the transparency of truth. Does that mean that we need transparency after all? If our goal is simply to find, for every predicate $\varphi$, *some* sentence that expresses the infinite disjunction of the $\varphi$s in the sense of (c), the answer is 'No'. For we can find a sentence different from $\exists x(\varphi(x) \wedge Tx)$ that does the job. Let us start again with the infinite disjunction $\bigvee\{\varphi(\ulcorner\psi\urcorner) \wedge \psi : \psi \in \mathcal{L}\}$. By assumption (b), its negation is equivalent to $\bigwedge\{\varphi(\ulcorner\psi\urcorner) \rightarrow \neg\psi : \psi \in \mathcal{L}\}$. By a small modification of our earlier argument (Proposition 4), this infinite conjunction has the same $T$-free consequences as the sentence $\forall x(\varphi(x) \rightarrow T\dot{\neg}x)$, and is therefore expressed by it, according to (a). Hence, by (c), the negation of that sentence, i.e. $\neg\forall x(\varphi(x) \rightarrow T\dot{\neg}x)$, expresses the infinite disjunction of the $\varphi$s. Moreover, note that our candidate sentence is indeed derivable from $\varphi(\ulcorner\psi\urcorner) \wedge \psi$ (given an elimination principle). Thus, if the finite-axiomatisation account is correct, it does not support the adoption of introduction principles for infinite disjunctions either.

Now, we do not necessarily want to suggest that the above account gives the correct explication of infinite disjunctions, and therefore more needs to be said about this. However, our focus in this paper is on the infinite-conjunction function. The above should serve to illustrate that even the case of infinite disjunctions does not necessarily entail the need for introduction principles.

## 6 Deflationism and the Logical Function of Truth

We have pointed out that *if* one's project is to formulate a formal theory of truth for the sole purpose of having a device that fulfils the quasi-logical functions discussed above, then what principles govern the use of the truth predicate in natural language is largely irrelevant. What matters is that the truth predicate is governed by principles that allow it to perform this function. While we haven't reached a final verdict in the previous sections, we hope we have succeeded in challenging the idea that full transparency is necessary for that. However, it is important to note that even if transparency principles turned out to play a major role, there are some important lessons to be drawn concerning deflationism about truth.

Deflationism takes the transparency of truth as its starting point. From it, deflationists extract mainly two conclusions, a negative and a positive one. The negative thesis is that truth is conceptually redundant and therefore cannot play a substantive role in philosophical discourse. The positive thesis is that it can play the quasi-logical roles that we have talked about in this paper. In fact, this is the sole reason why we have a truth predicate in our language in the first place, according to deflationism.

As a consequence, what truth principles deflationists should adopt for their *formal* theory of truth, or for their proposed regimentation of natural language, should depend *entirely* on their ability to ensure this function. If weaker, stronger or

simply just principles other than transparency turn out to be necessary and sufficient to guarantee these functions in a (semi-)formal setting, despite their starting point deflationists have no reasons to adopt (mere) disquotational principles in their formal theories. When it comes to devising a formal theory of truth or proposing a regimentation of our use of the truth predicate, the deflationist has no conceptual commitment to (mere) transparency.

However, one often finds a focus on disquotation in the literature on deflationism. To take just one example, Beall and Armour-Garb (2005) describe the T-biconditionals as "fundamental", "brute", "analytic", "necessary", "a priori", and "explanatorily and conceptually basic", while the function is only mentioned as an explanation why a predicate governed by the T-schema is not dispensable. This focus often translates into the search for formal systems that can accommodate transparency principles without leading to triviality, as mentioned before. But this gets things the wrong way round. If deflationism is right and the quasi-logical function is the sole reason for the existence of the truth predicate in the language, then the T-biconditionals are not "explanatorily and conceptually basic". The function of truth is basic and comes first, while the T-biconditionals are only relevant if and insofar as they enable the truth predicate to perform that function.[21]

What we would like to suggest is that the truth predicate is comparable to a tool. Coming up with a formal theory of truth is basically an engineering problem; it requires no conceptual analysis in the traditional sense. There could be several satisfactory theories of truth and choosing between them would only depend on how user-friendly and how effective the theories are. In the end, it might also turn out that truth simply cannot fulfil its intended function *unrestrictedly*, because of the paradoxes. This would be disappointing, and place limitations on the expressive power of our language. But it would not undermine the deflationists' account of truth. It is perfectly conceivable that a tool does not live up to all of our expectations.

One may worry that this reconceptualisation of deflationism deprives deflationism of the reason that made it so attractive for philosophers, namely, that it deflates the metaphysical puzzle of the nature of truth. But this worry is ungrounded. The anti-metaphysical stance of deflationism, we believe, does not rest on the idea that the notion of truth is governed by the T-biconditionals (a claim that may also be endorsed by non-deflationists), nor does it rest on the claim that the T-biconditionals are all there is to truth. Instead, we believe that the anti-metaphysical stance of deflationism derives (or is still derivable) from the idea that the notion of truth is a quasi-logical device, whose behaviour is governed by axioms or rules that allow it to perform its intended functions, and that *that* is all there is to truth. On this picture, truth is (roughly) on a par with logical connectives like negation or disjunction, and no more mysterious than them. It may very well be the case that the way in which the quasi-logical function of truth is implemented in natural language is in fact via transparency principles. (And it is a very interesting and difficult question how we

[21] This is not to say that truth is absolutely indispensable. It is as long as we remain within the limits of first-order logic, but we may also do with e.g. substitutional quantifiers. A similar point is made in Azzouni (2005). However, he draws rather different conclusions from it.

operate in natural language with a notion that, assuming it actually is governed by transparency principles, is classically inconsistent.) But this does not mean that we are committed to transparency principles when we devise a formal language or propose a regimentation of natural language that is suitable for scientific, mathematical or philosophical discourse. Of course, deflationists could be interested in formal theories of truth that capture the use of the truth predicate in natural language, and such a theory would presumably include transparency principles. However, it is by no means clear that such systems will turn out to be the ones that are most useful in scientific, mathematical or philosophical discourse.

This understanding has profound consequences on the deflationists' formal notion of truth. Interestingly, it helps the deflationist to deal with several objections that can be found in the literature regarding semantic paradoxes. Many philosophers think that the semantic paradoxes pose a special problem for deflationists,[22] presumably because they think that deflationists are conceptually committed to the T-schema when adopting a formal truth system. We will argue, on the contrary, that the deflationists are in a much *better* position to deal with the paradoxes than their opponents. (Thus, this is a strengthening of a claim made by Gupta (2005) to the effect that the paradoxes do not pose a special problem for deflationism.)

The T-schema leads to trivilality in classical logic, given minimal syntactic background assumptions. This forces us to reject certain instances of the T-schema or to reject certain classically valid (meta-)rules of inference (in addition, perhaps, to using a new conditional in the formulation of the T-biconditionals). It is sometimes said that such moves change the meaning of the concept of truth (cf. Field 2008, pp. 14–17). This might be turned into an argument against a specific truth theorist if her goal is to capture the truth predicate as it occurs in natural language. But the meaning of the truth predicate in natural language is not what deflationists are in general trying to capture in their formal theories of truth. What the deflationist is aiming for is merely a quasi-logical tool that fulfils certain expressive functions.

Thus the deflationist is clearly entitled to revise the naïve understanding of truth to formulate her formal systems. What restrictions should we impose on such a revision? Several authors have proposed different desiderata on a satisfactory theory of truth (cf. Halbach and Horsten 2005; Leitgeb 2007; Sheard 2002). We believe such lists often involve criteria which are only imported because truth seems to satisfy these criteria in natural language, without paying attention as to whether they play an important role in the expression of infinite conjunctions and disjunctions.

A desideratum that features prominently on such lists is that the theory "must satisfy a requirement of naturalness and simplicity. It must contain as few ad hoc elements as possible." (Halbach and Horsten 2005, p. 207). Given the paradoxes, this is often not the case: to solve them truth theorists usually turn to more complex principles than plain disquotation. Now, a solution is *ad hoc* if it's designed for a specific task or problem, lacking independent motivation. But if deflationism is right and the truth predicate exists solely for a certain expressive purpose, what other

[22] As witnessed by the fact that there is an entire collection devoted to that problem: Beall and Armour-Garb (2005).

motivation could there be than constructing the truth theory in such a way that it serves these logical needs? Whether the theory is artificial is largely irrelevant.[23]

Consider for example Tarski's hierarchy of truth predicates $T_1, T_2, T_3, \ldots$ Many philosophers have argued against Tarski's solution on the grounds that stratification is *ad hoc* and only motivated by the desire to evade the paradoxes; that truth is a univocal concept that is fragmented in the Tarskian approach; that typing (indexing) is not found in natural language. From the deflationist perspective discussed here, such objections have little force. The problem with the Tarskian solution, then, is not that it is artificial or *ad hoc*; the problem is that typed truth predicates simply do not satisfy our needs (entirely). For example, the hierarchy of truth predicates does not offer a means to express the soundness of the theory by a single sentence, or to express all instances of certain schemata at once. Whenever a formula $\varphi(x)$ defines an infinite set of sentences whose indices are unbound in rank, there is no way of axiomatising all the $\varphi$s by a single sentence using one of the truth predicates $T_\alpha$.[24]

Another requirement usually imposed on formal truth theories is that their so-called 'outer' and 'inner' logics coincide (cf. Halbach and Horsten 2005; Leitgeb 2007). The inner logic of a truth system is the set of sentences the theory can prove to be true, while the outer logic is simply the set of its theorems. In this paper we have suggested that T-Out should be part of any reasonable theory of truth. As it turns out, in classical logic this principle directly entails a discrepancy between the inner and the outer logic of a system. Let $\lambda$ be a liar sentence, as before. By T-Out we have

$$T\ulcorner\lambda\urcorner \rightarrow \lambda$$

But by the definition of $\lambda$, we also have

$$\neg T\ulcorner\lambda\urcorner \rightarrow \lambda$$

Therefore, $\lambda$. But then by the definition of $\lambda$, we have that $\neg T\ulcorner\lambda\urcorner$ as well and, hence, the conjunction $\lambda \wedge \neg T\ulcorner\lambda\urcorner$ becomes provable in the system.

But why should the outer and the inner logic of a truth theory coincide? Does this desideratum have a deflationist justification? Field (2008, chap. 6) claims that asserting $\lambda \wedge \neg T\ulcorner\lambda\urcorner$ violates a coherence principle: namely, that it is incoherent to assert a sentence and simultaneously asserting that that sentence is not true. Similarly, Horsten (2011, p. 127) claims that proving a sentence that is untrue is "a sure mark of philosophical unsoundness." Bacon (2015) poses a similar objection as a revenge problem for disquotational systems. Glanzberg (2005) has made analogous claims.

But why would it be incoherent to assert a sentence and simultaneously assert that it is not true? Presumably, it would be incoherent if we conceived of $\varphi$ and $T\ulcorner\varphi\urcorner$ as having the same meaning, or as being materially equivalent, or as having the

[23] Of course, there could be *practical* reasons why a 'natural' theory is preferable over a more 'artificial' one: a natural theory might be more user-friendly insofar as the rules of such a theory resemble the rules that the agent is used to in natural language.

[24] This does not necessarily mean that a type-free solution is to be preferred. It is not clear that the advantages of a type-free approach outweight the problems caused by the semantic paradoxes.

same semantic value, or maybe if we thought that a sentence is true if it holds in a model, or if what it says is the case. But, as we have argued, none of this needs to be part of the deflationists' formal notion to truth.

From the deflationist's point of view, the liar is best viewed not as a paradox but (if at all) as a limitation on the truth predicate's ability to fulfil its intended function. In this respect deflationists are in a more comfortable position than their opponents. For example, correspondence theorists have to explain how to accommodate a sentence that says of itself that it is untrue, given that a sentence is true (on their picture) if and only if what it says is the case. To put it pointedly: The paradoxes pose a bigger threat to correspondence theorists because the T-schema seems to be an essential part of their notion of truth, even if they think that there is *more* to truth than the T-biconditionals. Deflationists have no such commitment. They could just let some instances go and live with the fact that their truth predicate does not serve its intended role unrestrictedly. The correspondence theorists' theory of truth aims to unravel the nature of truth and therefore has to be sound; for the deflationist, the question of the soundness of the truth theory does not even arise. The deflationist's formal theory (or theories) of truth does not describe a property in the world but simply provides us with a quasi-logical tool, engineered to perform certain expressive functions. The question for the deflationist is simply whether, for instance, asserting untrue sentences substantially impairs these functions.

## 7 Agreement and Disagreement in Classical T-Out Theories

There is one reason one might worry that asserting that one's theorems are untrue deprives the truth predicate of its utility. The notion of truth is supposed to enable us to express agreement and disagreement with theories that cannot be finitely axiomatised (except by using a truth-like predicate, of course). Thus, the usual way of expressing agreement with a theory $\Gamma$ is to say "All theorems of $\Gamma$ are true". In the mathematical literature, this is also known as the *global reflection principle* for $\Gamma$ (see Kreisel and Levy 1968), formally

$$\forall x(Bew_{\Gamma}(x) \rightarrow Tx) \qquad (\mathrm{GRP}_{\Gamma})$$

where $\Gamma$ is a recursively enumerable (r.e.) theory and $Bew_{\Gamma}(x)$ is a predicate that weakly represents provability in $\Gamma$. Now the problem is that any classical theory that has 'incoherent' consequences—e.g. $\lambda \wedge \neg T\ulcorner\lambda\urcorner$—will be *inconsistent* with its own global reflection principle, and therefore seems to deprive us of the possibility to express agreement with our own theory. For simplicity, let us assume that our syntax theory is Peano arithmetic (PA).

**Proposition 8** *Assume $\Gamma$ is a classical r.e. theory extending* PA *and let $\lambda$ be a liar sentence. If $\Gamma \vdash \lambda$, then $\Gamma$+GRP $_{\Gamma}$ is inconsistent.*

*Proof* Assume $\Gamma \vdash \lambda$. Therefore, by definition of the liar, $\Gamma \vdash \neg T\ulcorner\lambda\urcorner$. Since $\Gamma$ is r.e. and extends PA we have $\Gamma \vdash Bew_{\Gamma}(\ulcorner\lambda\urcorner)$ and by $\mathrm{GRP}_{\Gamma}$ it follows that $\Gamma+\mathrm{GRP}_{\Gamma} \vdash T\ulcorner\lambda\urcorner$. Thus $\Gamma+\mathrm{GRP}_{\Gamma}$ is inconsistent. □

In particular, any classical theory extending PA + T-Out is inconsistent with its own global reflection principle. Field (2008) notices that T-Out-theorists might not only have problems with expressing agreement but also with expressing disagreement.

> [Assume that Jones] puts forward a quite elaborate gap theory involving T-Out. And suppose that I disagree with this theory overall, but can't quite decide which specific claims of the theory are problematic. It is natural for me to express my disagreement by saying "Not everything in Jones' theory is true". But this doesn't serve its purpose: since Jones himself, as a gap theorist, believes that important parts of his own theory aren't true, I haven't succeeded in expressing disagreement.
>
> Alternatively, suppose that Jones himself thinks that Brown's theory is wrong, but isn't quite sure which claims of it are wrong. Then he certainly can't express his disagreement by saying "Not everything in Brown's theory is true", since by his own lights that doesn't differentiate Brown's theory from his own. (Field 2008, p. 140)

We do not find these arguments very compelling. To begin with, for two people to agree or disagree on something, they first have to share the meaning of the words they will use to agree and disagree; in this case, the truth predicate. Presumably, Field's disagreement with Jones' theory isn't expressed in Jones' idiolect and, therefore, it isn't successful. Little has this to do with Jones' adoption of T-Out. Secondly, nothing forces Jones to express his agreement by saying "Everything in Browns' theory is true" or his disagreement by saying "Not everything in Brown's theory is true". This might be the way that natural language has chosen, but nothing prevents us from regimenting that use. If Jones disagrees with Brown's theory, Jones will suspect that there is some sentence $\varphi$ such that $\varphi$ is part of Brown's theory but $\neg\varphi$. And he can express that by saying "Something in Brown's theory is false" (where "$\varphi$ is false" is defined as "$\neg\varphi$ is true")—and *that* claim does differentiate Brown's theory from Jones'. Similarly, Jones can express agreement with his own theory by saying "Nothing in my theory is false" or "Everything in my theory is non-false", formally:

$$\forall x(Bew_{\Gamma}(x) \rightarrow \neg T\dot{\neg}x) \qquad (\mathrm{GRP}^{*}_{\Gamma})$$

Let us call $\mathrm{GRP}^{*}_{\Gamma}$ the 'modified global reflection principle' for $\Gamma$. The following shows that the modified global reflection principle can be consistently added to any T-Out-theory constructed over PA that has a standard model.

**Proposition 9** *If $\Gamma \supseteq$ PA contains T-Out and has a standard model (i.e., an $\omega$-model), then $\Gamma + GRP^{*}_{\Gamma}$ has a standard model too.*

*Proof* Assume otherwise. Then there must be a standard model $\mathcal{M}$ such that $\mathcal{M} \models \Gamma$ and $\mathcal{M} \models \exists x(Bew_{\Gamma}(x) \wedge T\dot{\neg}x)$. Since $\mathcal{M}$ is standard, there must be a sentence $\varphi$ such that $\mathcal{M} \models Bew_{\Gamma}(\ulcorner\varphi\urcorner) \wedge T\ulcorner\neg\varphi\urcorner$. By T-Out, $\mathcal{M} \models \neg\varphi$. But since $\mathcal{M}$

is standard, $\mathcal{M} \models Bew_\Gamma(\ulcorner\varphi\urcorner)$ implies $\Gamma \vdash \varphi$ and therefore $\mathcal{M} \models \varphi$. This contradicts the claim that $\mathcal{M} \models \neg\varphi$. □

One attractive feature of the ordinary global reflection principle is that it implies (given minimal assumptions) the consistency of the system in question. The modified reflection principle does the same job.

**Proposition 10** *If* $\Gamma \subseteq$ PA *contains* T-Out *and* $\Gamma \vdash T\ulcorner 0 \neq 1\urcorner$, *then* $\Gamma{+}GRP^*_\Gamma \vdash Con(\Gamma)$.[25]

*Proof* By Universal Instantiation, $\Gamma{+}\mathrm{GRP}^*_\Gamma \vdash Bew_\Gamma(\ulcorner 0 = 1\urcorner) \rightarrow \neg T\ulcorner 0 \neq 1\urcorner$. But since $\Gamma \vdash T\ulcorner 0 \neq 1\urcorner$, it follows that $\Gamma{+}\mathrm{GRP}^*_\Gamma \vdash \neg Bew_\Gamma(\ulcorner 0 = 1\urcorner)$. □

## 8 Conclusions

In this paper we have considered two accounts of what it means for general truth ascriptions of the form (InfC) to express the sentences satisfying the predicate $\varphi(x)$. According to one of them, (InfC) and the set of conditionals $\varphi(\ulcorner\psi\urcorner) \rightarrow \psi$ must be equivalent in some suitable sense. According to the other (which is the only formally worked out account in the literature), (InfC) must finitely axiomatise the the conditionals $\varphi(\ulcorner\psi\urcorner) \rightarrow \psi$ (relative to the truth theory). We have argued that the first account is flawed, and that the second one—if correct at all—justifies at best the adoption of a principle much weaker than transparency, namely, elimination principles for truth.

In Sect. 3, we have shown that, given reasonable background assumptions, elimination principles ensure what we have called the 'elimination property', which allows us (in some sense) to 'finitely axiomatise' infinite sets of sentences; so a full transparency principle is not needed for *that* purpose. Moreover, as we pointed out, in some cases full transparency does not even suffice to ensure the elimination property if the underlying logic is too weak.

In Sects. 4 and 5 it was also shown that, with the aid of some regimentation, the elimination property is sufficient to deal with a large number of generalisations that many authors believed to require full transparency. Thus, we conclude that the arguments offered in the literature so far do not provide enough support for the thesis that full transparency is needed for the logical function of truth. The upshot is that, first, we need a more thorough investigation of the expressive functions of the truth predicate, and of the principles each of them involves; and, second, that there are no conclusive reasons yet to abandon classical logic, contrary to what is usually believed. The move to non-classical logics is still in need of a supporting argument.

None of what we have said should be taken to imply that T-Out alone is sufficient as a theory of truth. On the one hand, T-Out alone does not enable us to prove any non-trivial generalisations about truth. On the other hand, at the end of Sect. 4 we

[25] Note that this defence of truth theories containing T-Out does not imply that the only truth principle we should adopt in our formal truth systems is T-Out. We might very well adopt in addition a restricted version of T-In, compositional principles, or whatever we need for our technical purposes. Some of these additional principles might in turn allow for a proof of $T\ulcorner 0 \neq 1\urcorner$.

have seen that T-Out does not account for *all* the inferences that we may want to make with generalisations. An answer to the question what other axioms are needed can only come from a full understanding of what it means for (InfC) to express all the $\varphi$s. We believe a promising line of thought is to understand truth as a means to emulate propositional (second-order, substitutional) quantification in a first-order setting. However, making this idea precise is by no means trivial. The main problem is to specify what 'emulation' comes to here, and there are several options to choose from. Working this out in detail is the goal of our next project.

In the final sections of the paper, we have argued that deflationism, properly understood, is not committed to many of the requirements that are usually imposed on truth systems, and immune to most of the objections frequently raised against them. Most of these requirements and objections rest on an understanding of the notion of truth that stems from its use in natural language or some correspondence intuition. Deflationists don't have to conform to this understanding when devising formal theories of truth. They only need to ensure (as far as possible) that their notion of truth can perform its intended expressive function. As a consequence, deflationists are in a much better position to deal with the paradoxes than their 'substantialist' opponents.

**Acknowledgements** This paper was presented to audiences in Buenos Aires, Canterbury, Ghent, Helsinki, Munich, New Jersey, and Tilburg; we thank the attendees of these talks for their valuable feedback. For their help in preparing this paper, we would like to thank JC Beall, Filip Buekens, Tim Button, Roy Cook, Hartry Field, Volker Halbach, Hannes Leitgeb, Dave Ripley, the Buenos Aires Logic Group, the Munich Center for Mathematical Philosophy, and an anonymous referee. This work was supported by the Alexander von Humboldt Foundation and the German Research Foundation (DFG, "Reference patterns of paradox").



## Appendix

In this appendix we give proofs of Observation 7, and similar limitative results for Priest's theory of truth and Kripke's fixed-point construction based on the Weak Kleene scheme that we mentioned at the end of Sect. 3. Before each proof we briefly introduce the corresponding system.

Recall $\mathcal{L}$ is our base, truth-free language, and $\mathcal{L}_T$ extends $\mathcal{L}$ with the monadic predicate symbol $T$. $\mathcal{L}$ contains enough vocabulary to talk about the syntax of $\mathcal{L}_T$. We assume $\mathcal{L}$'s standard model is classical and contains countably many objects in its domain. This allows us to work with substitutional semantics, and to keep it simple.

We first introduce the logic of paradox. LP is given by a family of models of $\mathcal{L}_T$ and a valuation scheme $v_{\text{LP}}$. An LP-model $\mathcal{M}$ behaves exactly like a classical model except that it assigns not only an extension $R^{\mathcal{M}^+}$ but also an anti-extension $R^{\mathcal{M}^-}$ to each $n$-adic predicate $R$ of $\mathcal{L}_T$, such that $R^{\mathcal{M}^+} \cup R^{\mathcal{M}^-} = |\mathcal{M}|^n$. $R^{\mathcal{M}^+} \cap R^{\mathcal{M}^-}$ may be non-empty. For each model $\mathcal{M}$, $v_{\text{LP}}^{\mathcal{M}}$ maps sentences of $\mathcal{L}_T$ into the set $\{0, \frac{1}{2}, 1\}$ as follows:

- $v_{\text{LP}}^{\mathcal{M}}(Rt_1 \ldots t_n) = \begin{cases} 1 & \text{if } \langle t_1, \ldots, t_n \rangle \in R^{\mathcal{M}^+} - R^{\mathcal{M}^-} \\ 0 & \text{if } \langle t_1, \ldots, t_n \rangle \in R^{\mathcal{M}^-} - R^{\mathcal{M}^+} \\ \frac{1}{2} & \text{otherwise} \end{cases}$
- $v_{\text{LP}}^{\mathcal{M}}(\neg\varphi) = 1 - v_{\text{LP}}^{\mathcal{M}}(\varphi)$
- $v_{\text{LP}}^{\mathcal{M}}(\varphi \wedge \psi) = min(v_{\text{LP}}^{\mathcal{M}}(\varphi), v_{\text{LP}}^{\mathcal{M}}(\psi))$
- $v_{\text{LP}}^{\mathcal{M}}(\varphi \rightarrow \psi) = max(1 - v_{\text{LP}}^{\mathcal{M}}(\varphi), v_{\text{LP}}^{\mathcal{M}}(\psi))$
- $v_{\text{LP}}^{\mathcal{M}}(\forall x\varphi) = min\{v_{\text{LP}}^{\mathcal{M}}(\varphi[t/x]) : t \text{ is a term}\}$

The intended reading of the truth values is *true* for 1, *false* for 0, and *both true and false* for $\frac{1}{2}$. Then, a sentence is true in an LP-model if the valuation based on that model assigns either value 1 or $\frac{1}{2}$ to it. Logical consequence is defined as truth preservation, as usual.

*Proof of Observation 7* Let $\Sigma \subseteq \mathcal{L}_T$ be a syntax theory for $\mathcal{L}_T$ (sound with respect to $\mathcal{L}$'s standard interpretation) formulated over a sound and complete calculus for LP, and let $\Gamma$ extend $\Sigma$ with the rules T-Intro and T-Elim (and, therefore, all instances of the T-schema). $\Gamma$ is non-trivial, for the standard model of $\mathcal{L}$ can be extended to an LP-model of $\Gamma$ (cf. Beall 2009). We assume $\Gamma$ contains a liar sentence $\lambda$, i.e. such that

$$\lambda \leftrightarrow \neg T\ulcorner\lambda\urcorner \tag{10}$$

We show the following instance of the elimination property does not hold in $\Gamma$:

$$\forall x(T(x \dot{\leftrightarrow} \ulcorner\lambda\urcorner) \rightarrow Tx), T(\ulcorner\bot\urcorner \dot{\leftrightarrow} \ulcorner\lambda\urcorner) \vdash \bot \tag{11}$$

where $\bot$ is a sentence that gets value 0 in all LP-models, and $\dot{\leftrightarrow}$ denotes the function that maps (the code of) two formulae $\varphi, \psi$ to (the code of) $\varphi \leftrightarrow \psi$. Let $\mathcal{M}$ be an LP-model of $\Gamma$ extending $\mathcal{L}$'s standard interpretation. Thus, the only predicate of $\mathcal{L}_T$ that may receive overlapping extension and anti-extension is $T$.

Note that, due to transparency and (10), $v_{\text{LP}}^{\mathcal{M}}(\lambda) = \frac{1}{2}$, and that, due to the clauses of LP's valuation scheme, every time one side of a biconditional gets value $\frac{1}{2}$, the value of the biconditional is also $\frac{1}{2}$. Thus, for every sentence $\varphi$, $v_{\text{LP}}^{\mathcal{M}}(\varphi \leftrightarrow \lambda) = \frac{1}{2}$. By transparency, $v_{\text{LP}}^{\mathcal{M}}(T\ulcorner\varphi \leftrightarrow \lambda\urcorner) = \frac{1}{2}$, and since $\mathcal{M}$ extends $\mathcal{L}$'s standard interpretation, $v_{\text{LP}}^{\mathcal{M}}(T(\ulcorner\varphi\urcorner \dot{\leftrightarrow} \ulcorner\lambda\urcorner)) = \frac{1}{2}$ too, which means the second premise in (11) is true in $\mathcal{M}$. This also implies that $v_{\text{LP}}^{\mathcal{M}}(T(\ulcorner\varphi\urcorner \dot{\leftrightarrow} \ulcorner\lambda\urcorner) \rightarrow T\ulcorner\varphi\urcorner) \geqslant \frac{1}{2}$. Under the reasonable

assumption that only (standard) sentences belong to $T^{\mathcal{M}^+}$, we also get that $v_{\text{LP}}^{\mathcal{M}}(\forall x(T(x \dot{\leftrightarrow} \ulcorner\lambda\urcorner) \rightarrow Tx)) \geqslant \frac{1}{2}$, this is, the first premise in (11) is true as well. But $v_{\text{LP}}^{\mathcal{M}}(\bot) = 0$. □

We now show how Priest's disquotational theory does not validate the contrapositive of the elimination property, that is:

$$\varphi(\ulcorner\psi\urcorner), \neg\psi \vdash \neg\forall x(\varphi(x) \rightarrow Tx) \tag{12}$$

This theory isn't explicitly formulated but can be extracted from Priest (2006). There he fully endorses the T-schema and argues that it must hold without restriction for the sake of the expressive function of truth (cf. Priest 2006, chap. 4). Priest works over LP, but in order to avoid the problems stated in Observation 7—viz. the failure of Modus Ponens—he supplements the logic with a new, non-extensional conditional with which he formulates his version of the T-schema. Furthermore, this new conditional, he argues, must be non-contraposible, i.e. $\varphi \rightarrow \psi$ should not necessarily imply $\neg\psi \rightarrow \neg\varphi$. In his own words, "There seems to be no reason why, in general, if $\varphi$ is a dialetheia, $T\ulcorner\varphi\urcorner$ is too. If $\varphi$ is a dialetheia, $T\ulcorner\varphi\urcorner$ is certainly true, but it might be simply true, and not also false" (Priest 2006, p. 79). We call the resulting logic 'PL', for 'Priest's Logic'.

PL is given by a family of models of $\mathcal{L}_T$ and a valuation scheme $v_{\text{PL}}$. A PL-model $\mathcal{M}$ consists of a set of possible worlds $W_{\mathcal{M}}$, a binary surjective relation $R_{\mathcal{M}} \subseteq W_{\mathcal{M}}^2$, and a function assigning an LP-model $\mathcal{M}_w$ to each $w \in W_{\mathcal{M}}$, all with the same domain. $v_{\text{PL}}$ behaves exactly like $v_{\text{LP}}$ for the extensional connectives, now relativized to each world $w \in W_{\mathcal{M}}$. The conditional, instead, is defined by the following clause:

- $v_{\text{PL}}^{w}(\varphi \rightarrow \psi) \geqslant \frac{1}{2}$ iff, for all $w' R_{\mathcal{M}} w$, if $v_{\text{PL}}^{w'}(\varphi) \geqslant \frac{1}{2}$, then $v_{\text{PL}}^{w'}(\psi) \geqslant \frac{1}{2}$

A sentence $\varphi$ is true in $\mathcal{M}$ if and only if $v_{\text{PL}}^{w}(\varphi) \geqslant \frac{1}{2}$ in every world $w \in W_{\mathcal{M}}$, and logical consequence is defined in the same way as LP's.

Let Priest's theory $\Gamma \subseteq \mathcal{L}_T$ be formulated over a calculus sound with respect to PL. $\Gamma$ consists of all instances of the T-schema (formulated with $\rightarrow$) plus a biconditional $\psi \Leftrightarrow \varphi(\ulcorner\psi\urcorner)$ for each open formula $\varphi(x) \in \mathcal{L}_T$, where $\Rightarrow$ expresses the material conditional of LP—i.e. $\varphi \Rightarrow \psi$ abbreviates $\neg(\varphi \wedge \neg\psi)$—and $\Leftrightarrow$ is defined in terms of $\Rightarrow$, as usual. Thus, the liar sentence $\lambda$ in $\Gamma$ is given by

$$\lambda \Leftrightarrow \neg T\ulcorner\lambda\urcorner \tag{13}$$

Assume $\Gamma$ has an PL-model $\mathcal{M}$ where identity behaves classically, such that at every world $w \in W_{\mathcal{M}}$ $\lambda$ gets value $\frac{1}{2}$ while $T\ulcorner\lambda\urcorner$ gets value 1, as Priest wants. This assumption seems plausible: if $\lambda$ gets $\frac{1}{2}$ and $T\ulcorner\lambda\urcorner$ gets 1 at each world $w$, then $v_{\text{PL}}^{w}(\lambda \leftrightarrow T\ulcorner\lambda\urcorner) \geqslant \frac{1}{2}$, by the clause of $v_{\text{PL}}$ for the conditional. But also, by the clause for negation, $v_{\text{PL}}^{w}(\neg T\ulcorner\lambda\urcorner) = 0$, which means that $v_{\text{PL}}^{w}(\lambda \Leftrightarrow \neg T\ulcorner\lambda\urcorner) = \frac{1}{2}$.[26] We show $\mathcal{M}$ is a counter-model of the following instance of (12):

[26] Note that, if instead of postulating (13) as an axiom $\lambda$ had been obtained by a standard diagonalisation process in a base theory of syntax, $\lambda$ and $\neg T\ulcorner\lambda\urcorner$ would have been forced to have the same truth value,

$$\ulcorner\lambda\urcorner = \ulcorner\lambda\urcorner, \neg\lambda \vdash \neg\forall x(x = \ulcorner\lambda\urcorner \rightarrow Tx) \tag{14}$$

By the clause for negation, we have that $v^w_{\text{PL}}(\neg\lambda) = \frac{1}{2}$ at every $w \in W_{\mathcal{M}}$. By the classicality of identity, we also have that $v^w_{\text{PL}}(\ulcorner\lambda\urcorner = \ulcorner\lambda\urcorner) = 1$. This gives us the truth of both premises in (14). However, the conclusion is false. Let $w' R_{\mathcal{M}} w$. While $v^{w'}_{\text{PL}}(\ulcorner\lambda\urcorner = \ulcorner\lambda\urcorner) = v^{w'}_{\text{PL}}(T\ulcorner\lambda\urcorner) = 1$, $v^{w'}_{\text{PL}}(t = \ulcorner\lambda\urcorner) = 0$ for each term $t$ such that $\mathcal{M}_{w'} \nvDash t \neq \ulcorner\lambda\urcorner$. Thus, by the clauses for the conditional and the universal quantifier, $v^w_{\text{PL}}(\forall x(x = \ulcorner\lambda\urcorner \rightarrow Tx)) = 1$ and, therefore, $v^w_{\text{PL}}(\neg\forall x(x = \ulcorner\lambda\urcorner \rightarrow Tx)) = 0$.

Next we prove Kripke's fixed-point theory Kripke (1975) with the Weak Kleene scheme also fails to satisfy (12). First, we introduce Kleene's weak three-valued logic B3. B3 is given by a family of models of $\mathcal{L}_T$ and a valuation scheme $v_{\text{WK}}$. B3-models $\mathcal{M}$ also assign an extension $R^{\mathcal{M}^+}$ and an anti-extension $R^{\mathcal{M}^-}$ to each $n$-adic predicate $R$ of $\mathcal{L}_T$, but unlike LP-models, $R^{\mathcal{M}^+} \cap R^{\mathcal{M}^-}$ must be empty and $R^{\mathcal{M}^+} \cup R^{\mathcal{M}^-}$ can be a proper subset of $|\mathcal{M}|^n$. For each B3-model $\mathcal{M}$, $v^{\mathcal{M}}_{\text{WK}}$ maps sentences of $\mathcal{L}_T$ intro de set $\{0, \frac{1}{2}, 1\}$ in the following way:

- $v^{\mathcal{M}}_{\text{WK}}(Rt_1 \ldots t_n) = \begin{cases} 1 & \text{if } \langle t_1, \ldots, t_n\rangle \in R^{\mathcal{M}^+} \\ 0 & \text{if } \langle t_1, \ldots, t_n\rangle \in R^{\mathcal{M}^-} \\ \frac{1}{2} & \text{otherwise} \end{cases}$
- $v^{\mathcal{M}}_{\text{WK}}(\neg\varphi) = 1 - v^{\mathcal{M}}_{\text{WK}}(\varphi)$
- $v^{\mathcal{M}}_{\text{WK}}(\varphi \wedge \psi) = min(v^{\mathcal{M}}_{\text{WK}}(\varphi), v^{\mathcal{M}}_{\text{WK}}(\psi))$
- $v^{\mathcal{M}}_{\text{WK}}(\varphi \rightarrow \psi) = 1 - min\{1 - v^{\mathcal{M}}_{\text{WK}}(\varphi), v^{\mathcal{M}}_{\text{WK}}(\psi)\}$
- $v^{\mathcal{M}}_{\text{WK}}(\forall x\varphi) = min\{v^{\mathcal{M}}_{\text{WK}}(\varphi[t/x]) : t \text{ is a term}\}$

where the ordering of the truth values is given by $\frac{1}{2} < 0 < 1$. While 1 and 0 keep their usual meaning, $\frac{1}{2}$ is better understood as *nonsense*. A nonsensical component renders the whole expression nonsensical, so every sentence containing a subsentence with value $\frac{1}{2}$ also gets value $\frac{1}{2}$. A sentence is true in a B3-model if the valuation based on that model gives it value 1, and logical consequence is defined as truth preservation.

Kripke's fixed-point theory based on B3 consists of a class $\Phi$ of B3-models $\mathcal{M}$ of $\mathcal{L}_T$ that extend $\mathcal{L}$'s standard interpretation in a way that, for every sentence $\varphi$, $v^{\mathcal{M}}_{\text{WK}}(T\ulcorner\varphi\urcorner) = v^{\mathcal{M}}_{\text{WK}}(\varphi)$, which ensures the transparency of the truth predicate. It can be shown that $\Phi$ is non-empty (cf. Martin and Woodruff 1975). We show the following instance of (12) does not hold in any model of $\Phi$:

$$\ulcorner 0 \neq 0\urcorner = \ulcorner 0 \neq 0\urcorner, \neg 0 \neq 0 \vdash \neg\forall x(x = \ulcorner 0 \neq 0\urcorner \rightarrow Tx) \tag{15}$$

Let $\mathcal{M} \in \Phi$. We assume $\mathcal{L}$ is rich enough to diagonalise every open formula of the language. Since $\mathcal{M}$ extends $\mathcal{L}$'s standard interpretation, there's a liar sentence $\lambda$

Footnote 26 continued
namely, $\frac{1}{2}$, which would have meant $T\ulcorner\lambda\urcorner$ also got value $\frac{1}{2}$. And if we replaced $\Leftrightarrow$ with the new, non-extensional biconditional $\leftrightarrow$, it wouldn't be possible to assign value $\frac{1}{2}$ to $\lambda$ and 0 to $\neg T\ulcorner\lambda\urcorner$ at every possible world. The same can be said of the use of inference rules.

such that (10) is true in $\mathcal{M}$. This implies that $v^{\mathcal{M}}_{\mathrm{WK}}(\neg T\ulcorner\lambda\urcorner) = v^{\mathcal{M}}_{\mathrm{WK}}(T\ulcorner\lambda\urcorner) = \frac{1}{2}$. Also, $v^{\mathcal{M}}_{\mathrm{WK}}(\neg 0 \neq 0) = v^{\mathcal{M}}_{\mathrm{WK}}(\ulcorner 0 \neq 0\urcorner = \ulcorner 0 \neq 0\urcorner) = 1$; so both premises in (15) are true in $\mathcal{M}$. The conclusion, however, isn't. To see it note that one of the instances of the universal statement $\forall x(x = \ulcorner 0 \neq 0\urcorner \rightarrow Tx)$ is the conditional $\ulcorner\lambda\urcorner = \ulcorner 0 \neq 0\urcorner \rightarrow T\ulcorner\lambda\urcorner$, whose truth value is given by $1 - min\{1 - v^{\mathcal{M}}_{\mathrm{WK}}(\ulcorner\lambda\urcorner = \ulcorner 0 \neq 0\urcorner), v^{\mathcal{M}}_{\mathrm{WK}}(T\ulcorner\lambda\urcorner)\}$, which is $\frac{1}{2}$, since $v^{\mathcal{M}}_{\mathrm{WK}}(T\ulcorner\lambda\urcorner) = \frac{1}{2}$. As a consequence, $\forall x(x = \ulcorner 0 \neq 0\urcorner \rightarrow Tx)$ gets value $\frac{1}{2}$, and so does $\neg\forall x(x = \ulcorner 0 \neq 0\urcorner \rightarrow Tx)$, by the clause of $v_{\mathrm{WK}}$ for the quantifier.

## References


Azzouni, J. (2005). Anaphorically restricted quantifiers and paradoxes. In J. C. Beall & J. Armour-Garb (Eds.), *Deflationism and paradox* (pp. 250–273). Oxford: Oxford University Press.

Bacon, A. (2015). Can the classical logician avoid the revenge paradoxes? *Philosophical Review*, *124*, 299–352.

Baker, S., Ireland, A., & Smaill, A. (1992). On the use of the constructive omega-rule within automated deduction. In A. Voronkov (Ed.), *Logic programming and automated reasoning. LPAR 1992. Lecture notes in computer science (lecture notes in artificial intelligence)* (Vol. 624). Berlin: Springer.

Beall, J. C. (2009). *Spandrels of truth*. Oxford: Oxford University Press.

Beall, J. C. (2013). Free of detachment: Logic, rationality, and gluts. *Noûs*, *49*, 410–423.

Beall, J. C., & Armour-Garb, B. (Eds.). (2005). *Deflationism and paradox*. New York: Oxford University Press.

Cobreros, P., Egré, P., Ripley, D., & van Rooij, R. (2013). Reaching transparent truth. *Mind*, *122*, 841–866.

Craig, W., & Vaught, R. (1958). Finite axiomatizability using additional predicates. *Journal of Symbolic Logic*, *23*, 289–308.

Field, H. (2007). Solving the paradoxes, escaping revenge. In J. C. Beall (Ed.), *Revenge of the liar* (pp. 78–144). Oxford: Oxford University Press.

Field, H. (2008). *Saving truth from paradox*. New York: Oxford University Press.

Fujimoto, K. (2012). Classes and truths in set theory. *Annals of Pure and Applied Logic*, *163*, 1484–1523.

Glanzberg, M. (2005). Minimalism, deflationism and paradoxes. In J. C. Beall & B. Armour-Garb (Eds.), *Deflationism and paradox* (pp. 107–132). Oxford: Oxford University Press.

Gupta, A. (1993). A critique of deflationism. *Philosophical Topics*, *21*, 57–81.

Gupta, A. (2005). Do the paradoxes pose a special problem for deflationism? In J. C. Beall & B. Armour-Garb (Eds.), *Deflationism and paradox* (pp. 133–147). Oxford: Oxford University Press.

Gupta, A., & Belnap, N. D. (1993). *The revision theory of truth*. Cambridge: MIT Press.

Halbach, V. (1999). Disquotationalism and infinite conjunctions. *Mind*, *108*, 1–22.

Halbach, V. (2011). *Axiomatic theories of truth*. Cambridge: Cambridge University Press.

Halbach, V. (2016). Prolegomena zu einer jeden künftigen definition von wissen, die als lösung des gettierproblems wird auftreten können. In W. Freitag, H. Rott, H. Sturm, & A. Zinke (Eds.), *Von Rang und Namen: Essays in honour of Wolfgang Spohn* (pp. 147–172). Paderborn: Mentis.

Halbach, V., & Horsten, L. (2005). The deflationist's axioms for truth. In B. Armour-Garb & J. C. Beall (Eds.), *Deflationism and paradox*. Oxford: Oxford University Press.

Heck, R, Jr. (2004). Truth and disquotation. *Synthese*, *142*, 317–352.

Horsten, L. (2011). *The Tarskian turn: Deflationism and axiomatic truth*. Cambridge: MIT Press.

Horwich, P. (1998). *Truth* (2nd ed.). Oxford: Blackwell.

Huemer, M. (2005). Logical properties of warrant. *Philosphical Studies*, *122*, 171–182.

Kemp, G. (2005). Disquotationalism and expressiveness. *Journal of Philosophical Logic*, *34*, 327–332.

Kreisel, G., & Levy, A. (1968). Reflection principles and their use for establishing the complexity of axiomatic systems. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, *14*, 97–142.

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

Leeds, S. (1978). Theories of reference and truth. *Erkenntnis*, *13*, 111–129.
Leitgeb, H. (2007). What theories of truth should be like (but cannot be). *Philosophy Compass*, *2*, 276–290.
Martin, R. L., & Woodruff, P. W. (1975). On representing 'True-in-L' in L. *Philosophia*, *5*, 213–217.
Priest, G. (2006). *In contradiction*. New York: Oxford University Press.
Priest, G. (2008). *An introduction to non-classical logic* (2nd ed.). Cambridge: Cambridge University Press.
Putnam, H. (1978). *Meaning and the moral sciences*. London: Routledge.
Quine, W. V. O. (1970). *Philosophy of logic*. Cambridge: Harvard University Press.
Raatikainen, P. (2005). On Horwich's way out. *Analysis*, *65*, 175–177.
Restall, G. (2000). *An introduction to substructural logics*. New York/London: Routledge.
Ripley, D. (2015). Comparing substructural theories of truth. *Ergo*, *2*, 13.
Sheard, M. (2002). Truth, provability and naive criteria. In V. Halbach & L. Horsten (Eds.), *Principles of truth* (pp. 169–181). Frankfurt am Main: Hänsel-Hohenhausen.
Tarski, A. (1935). The concept of truth in formalized languages. In J. Corcoran (Ed.), *Logic, Semantics, Metamathematics* (pp. 152–278). Oxford: Clarendon Press.