Inferential Deflationism

Forthcoming in *The Philosophical Review*

Luca Incurvati & Julian J. Schlöder

Abstract

Deflationists about truth hold that the function of the truth predicate is to enable us to make certain assertions we could not otherwise make. Pragmatists claim that the utility of negation lies in its role in registering incompatibility. The pragmatist insight about negation has been successfully incorporated into bilateral theories of content, which take the meaning of negation to be inferentially explained in terms of the speech act of rejection. We implement the deflationist insight in a bilateral theory by taking the meaning of the truth predicate to be explained by its inferential relation to assertion. We combine this account of the meaning of the truth predicate with a new diagnosis of the Liar Paradox: its derivation requires the truth rules to preserve evidence, but these rules only preserve commitment. The result is a novel inferential deflationist theory of truth, which solves the Liar Paradox in a principled manner. We end by showing that our theory and simple extensions thereof have the resources to axiomatize the internal logic of several supervaluational hierarchies, including Cantini's. This solves open problems of Halbach (2011) and Horsten (2011).

Keywords: truth; deflationism; inferentialism; pragmatism; supervaluationism

1 Truth, Negation and the Liar

Earlier versions of this material were presented at the Universities of Amsterdam, Buenos Aires, Connecticut at Storrs, Texas at Austin, and Tübingen, as well as the Graduate Center of the City Univer-

We may want to voice our agreement with Thomas even if we forgot what exactly he said. The truth predicate allows us to do that: we refer to Thomas's claim as *what Thomas said* and predicate truth of it. The truth predicate can be used for *indirect* endorsement. Similarly, although we cannot assert infinitely many sentences, we may want to endorse every instance of the law of excluded middle. The truth predicate makes this possible: we universally quantify over all sentences of the form *p or not p* and predicate truth of them.[1] The truth predicate can be used for *compendious* endorsement.[2]

According to deflationists, we have a *need* for a device of indirect or compendious endorsement. It is in response to this need, they claim, that the truth predicate has come to be part of our languages. The expressive *function* of the truth predicate is its *raison d'être*. Other mechanisms, such as substitutional quantification, could have realized this function, but our languages have evolved so that it is realized by the truth predicate (Horwich 1998, 124ff.).

This is an appealing account of the reason we have a truth predicate in our language, but it faces a challenge: the truth predicate may occur in embedded contexts in which its use cannot be taken to indicate endorsement. In uttering *if what Thomas said is true, we should leave*, one is not thereby endorsing what Thomas said (Soames 1999, 237; Picollo and Schindler 2018, 329). This is a version of the *Frege-Geach embedding problem*. Although the problem has mostly been discussed in connection with non-cognitivism in meta-ethics, it was initially raised by Gottlob Frege (1919) against a type of account

sity of New York and the Seminar of the London Group for Formal Philosophy. We are grateful to the members of those audiences for their valuable feedback. For comments and discussions, we would like to thank Tim Button, Julien Murzi, Carlo Nicolai, Lucas Rosenblatt, Thomas Schindler, Stewart Shapiro, and the referees and editors of this journal. This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 758540) within the project *From the Expression of Disagreement to New Foundations for Expressivist Semantics*.

[1] Truth can be predicated of many things, such as claims, sentences and propositions. For the purposes of this paper, we take *sentences* to be the primary truth bearers and will state the views to be discussed accordingly. The question of what the primary truth bearers are is a vexed one (Künne 2003, 563ff.), but we need not enter it here. The arguments of this paper can be easily recast by taking the primary truth bearers to be, for instance, propositions or claims.

[2] The terminology of 'indirect' and 'compendious' endorsement is Crispin Wright's (1992). The idea that the truth predicate is a device for indirect and compendious endorsement goes back at least as far as Quine 1970.

of negation that bears remarkable similarities to the deflationist account of truth. It is there, we suggest, that deflationists should look to develop their view so as to avoid the Frege-Geach problem.

Deflationists tell a functionalist story about the truth predicate. The truth predicate serves to fulfill an expressive need. The pragmatist Huw Price (1990) tells a similarly functionalist story about negation. Negation too serves to fulfil a certain expressive need: to register perceived incompatibilities. Negation allows us to do so by allowing us to *reject* claims. Similar accounts of negation have been suggested by other pragmatists, including Charles Sanders Peirce (1905), Frank Ramsey (1927), Wilfrid Sellars (1969) and Robert Brandom (1994).[3]

Like the functionalist account of truth, the functionalist account of negation faces the Frege-Geach problem: negation may occur in embedded contexts in which its use cannot be taken to indicate the performance of a rejection. For instance, one may sincerely utter *if Janet does not get the job, she cannot buy a house* even though one believes that it is likely that Janet *will* get the job. *Bilateral* theories of content afford the means to retain the key insight of the functionalist account of negation while avoiding the Frege-Geach problem. This insight is that negation is to be explained in terms of rejection. Bilateral theories hold that this explanation is *inferential*: the meaning of negation is given by its inferential relation to rejection (Smiley 1996; Rumfitt 2000; Incurvati and Schlöder 2017).[4]

We use the same strategy to rescue the functionalist account of truth from the Frege-Geach problem. The key insight of deflationism is that the truth predicate is to be explained in terms of its role in indirect or compendious endorsement. Within the bilateral framework, however, one may take this explanation to be inferential: the meaning of the truth predicate is given by its inferential relation to assertion, the speech act expressing endorsement. The result is a version of *inferential* deflationism, the view

[3] For an enlightening history of rejection in American pragmatism, see Beisecker 2019. On Cambridge pragmatism, including Ramsey and Price, see Misak 2016.

[4] In Incurvati and Schlöder 2021 we show that the bilateral solution also applies to more recent versions of the Frege-Geach problem (Schroeder 2008) and that it generalizes beyond the case of negation (in particular, encompassing the meta-ethical case).

that the meaning of the truth predicate is given by the truth rules allowing us to infer $\ulcorner p \urcorner$ *is true* from p and *vice versa* (where $\ulcorner p \urcorner$ is a name for p).[5] Inferential deflationism was considered and rejected by Anil Gupta (1993, 74–75) and later endorsed by Alan Weir (1996) and Leon Horsten (2009). We are the first to advocate it to vindicate the functionalist insight of deflationism whilst escaping the Frege-Geach problem.

But there is more. Once inferential deflationism is embedded within a bilateral framework, a novel and principled solution of the semantic paradoxes becomes available. Having argued for a functionalist account of truth, deflationists wield Occam's Razor to conclude that the right semantics for the truth predicate is the simplest one that accounts for its expressive function (Williams 1988). This function, deflationists continue, is captured by the truth rules. However, this immediately leads to a Liar paradox. The standard derivation goes as follows. Consider a sentence L equivalent to $\ulcorner L \urcorner$ *is not true*. Towards a *reductio*, assume L. From L, it follows that $\ulcorner L \urcorner$ *is true* by the truth rules and that $\ulcorner L \urcorner$ *is not true* by the definition of L. Contradiction. By *reductio*, we obtain *not* L. But from *not* L it follows that $\ulcorner L \urcorner$ *is true* by the definition of L, which entails L by the truth rules. Contradiction.

Deflationist opinion is divided on what the best course of action is in the wake of paradox. Paul Horwich's (1998) suggestion amounts to blaming the truth rules and restricting them to a certain subset from which no paradox follows. This move damages the simplicity of deflationism, perhaps disastrously (McGee 1992). Others take the culprit to be the classical laws of negation and recommend relinquishing them in favor of a paracomplete (Field 2008) or paraconsistent logic (Priest 1979; Beall 2009).

In this paper, we present a different, new diagnosis and solution of the Liar Paradox. What goes wrong in the standard derivation of the paradox is that the truth rules are used as if they preserved evidence. However, as we argue in Section 4, neither the inference from p to $\ulcorner p \urcorner$ *is true* nor the inference from $\ulcorner p \urcorner$ *is true* to p preserves evidence. Properly taking this into account blocks the derivation of the Liar Paradox. The very same solution also blocks the strong Liar paradox.

[5] To be precise: the expression 'p' is a meta-language name for an object-language sentence and '$\ulcorner p \urcorner$' is a meta-language name for an object-language name for the object-language sentence denoted by 'p'.

Although the truth rules do not preserve evidence, they do preserve commitment. These rules can therefore be taken to give the meaning of the truth predicate, as inferential deflationism maintains. This move is available to us only because in our preferred bilateral framework, inference rules need not preserve evidence, but only need to preserve commitment. By combining a bilateral account of the meaning of the truth predicate with our diagnosis of the semantic paradoxes we obtain a theory of truth which avoids the paradoxes and retains the classical laws of negation.

We develop the theory within a particular version of the bilateral framework, which distinguishes between *weak* and *strong* rejections (first developed in Incurvati and Schlöder 2017). The weak rejection of p, unlike its strong rejection, does not entail the negative assertion of p. One upshot is that in our theory one can—indeed must—reject both the Liar sentence and its negation.

But there is more. By applying our theory of truth to the collection of all true arithmetical sentences and extending it with the ω-Rule, we can establish a precise relation with the supervaluational hierarchy of Bas van Fraassen (1971). To wit, the asserted sentences that the extended theory deems to be true are exactly those that belong to the extension of the truth predicate defined by the van Fraassen hierarchy. Motivated extensions of our theory yield results for the truth predicates defined by Andrea Cantini's (1990) supervaluational hierarchy and the maximally consistent hierarchy first discussed by Saul Kripke (1975). Similar results have been obtained by Toby Meadows (2015) and Johannes Stern (2018), but we are the first to provide simple, natural deduction calculi for membership in all these truth predicates. Crucially, the structure of our theory allows us to separate an 'external' logic from the 'internal' characterization of the truth predicate, which results in natural recursive axiomatizations of these supervaluational theories of truth, solving open problems of Volker Halbach (2011) and Leon Horsten (2011).

We begin by presenting the standard bilateral framework in Section 2 and our preferred version of it in Section 3. We show how to extend the framework with a truth predicate in Section 4, where we also defend the idea that the truth rules do not preserve evidence, but commitment. Having shown how our theory of truth deals with the Liar Paradox, we consider some objections and the strong Liar paradox in Section 5. We go on to use the theory to provide axiomatizations of various supervaluational hierarchies in Section 6. We conclude in Section 7 by commenting on the relationship between deflationism and the supervaluational approach.

2 Bilateralism

Price (1990) argues that there must be a primitive operation of rejection, as without it we could not inform someone that our claims are incompatible with theirs. If someone asserts some *p* and we try to refute them by asserting a contrary *q*, we may fail because our interlocutor may not realize that *p* and *q* are incompatible. Even if our interlocutor understands the truth table for negation and we assert *not p*, Price continues, we may still fail if our interlocutor does not realize that truth and falsity are incompatible. Thus, to inform someone of a perceived incompatibility, we need a primitive operation to *register* this incompatibility. This operation, Price concludes, is rejection and the meaning of negation is given by its function to indicate rejection.

Price's functionalist account of negation has to contend with the Frege-Geach problem: when suitably embedded, negation cannot be taken to serve to perform a rejection. Bilateralism offers a way to solve the Frege-Geach problem whilst giving its due to the idea that the meaning of negation is explained in terms of rejection.

According to bilateralism, the meaning of an expression is given by conditions on the primitive speech acts of assertion and rejection. The meaning-conferring conditions are formulated by means of inference rules in a natural deduction system. Bilateralism is therefore a version of *inferentialist* semantics: the meaning of an expression is given by its role in inferences. Formally, *sentences* are obtained from a countable set of propositional atoms, conjunction $\wedge$ and negation $\neg$ in the usual way. We abbreviate $\neg(\neg A \wedge \neg B)$ as $A \vee B$ and $\neg(A \wedge \neg B)$ as $A \rightarrow B$. *Formulae* are sentences prefixed (or *signed*) with *force markers* for assertion and rejection. The formula $+A$ represents the assertion of the sentence A and $-A$ its rejection.

The meaning of negation can still be explained in terms of rejection, in keeping with the functionalist account. But rather than *indicating* rejection, negation is *inferentially explained* in terms of rejection. In particular, the rules for negation in standard bilateral systems (Smiley 1996; Rumfitt 2000) allow one to move from the rejection to negative assertion and *vice versa*, and from assertion to negative rejection and *vice versa*.

$$(+\neg\text{I.})\ \frac{-A}{+\neg A} \qquad (+\neg\text{E.})\ \frac{+\neg A}{-A} \qquad (-\neg\text{I.})\ \frac{+A}{-\neg A} \qquad (-\neg\text{E.})\ \frac{-\neg A}{+A}$$

As Ian Rumfitt (2000) notes, these rules satisfy the usual criteria on the admissibility of inference rules, notably harmony.

Negation has been explained in terms of rejection, but for the functionalist account to be vindicated, rejection needs to express incompatibility. This can be achieved by laying down rules ensuring that assertion and rejection are incompatible. The appropriate rules are (Rejection), which permits to infer absurdity from having asserted and rejected the same content, and the *Smileian reductio* rules, which state how to discharge an inferred absurdity (Smiley 1996).[6]

$$(\text{Rejection})\ \frac{+A \qquad -A}{\bot} \qquad (\text{SR}_1)\ \frac{\begin{array}{c}[+A]\\ \vdots\\ \bot\end{array}}{-A} \qquad (\text{SR}_2)\ \frac{\begin{array}{c}[-A]\\ \vdots\\ \bot\end{array}}{+A}$$

These rules are known as *coordination principles*, in that they do not characterize the meaning of an operator, but govern the interaction of assertion and rejection.

If negation is defined by the quartet of rules above and the interaction of assertion and rejection is governed by the coordination principles, *reductio* and double negation elimination are valid in the logic of assertion. The bilateralist has inferentially defined classical negation, contrary to the widespread view that associates inferentialism with intuitionistic logic (Dummett 1991).

We can now see how bilateralism deals with the Frege-Geach embedding problem. The bilateralist agrees with Frege (1919) that *not* cannot be taken to indicate rejection,

[6] In bilateralist logics, '$\bot$' is a punctuation sign indicating a logical dead end (Tennant 1999). It is therefore not prefixed with a force marker (Rumfitt 2000).

since when one utters *if Janet does not get the job, she cannot buy a house*, the first occurrence of *not* does not indicate rejection. But, the bilateralist notes, this utterance does commit the speaker to *Janet cannot buy a house* should they reject *Janet will get the job*. The inference rules giving the meaning of negation in terms of rejection ensure this. In particular, from a rejection of *Janet will get the job* it follows that *Janet will not get the job*. This and the original utterance jointly entail, by simple *modus ponens*, that Janet cannot buy a house, as desired. The bilateralist maintains that negative assertions license rejections in unembbeded contexts but recognizes Frege's point that this is not so in embedded contexts. By explaining the meaning of negation in terms of its inferential relation to rejection (as per the bilateral negation rules), the bilateralist can account for the behavior of negation in all contexts and keep its meaning constant across these contexts.

But if *not* does not indicate rejection, it may seem a simple act of faith to believe in the existence of a primitive speech act of rejection. Bilateralists such as Smiley (1996) and Rumfitt (2000) argue that to find examples of rejection in natural language we need to look no further than negative answers to self-posed polar questions: although *not* does not indicate rejection, *no* does. Similarly, *yes* indicates assertion. The resulting picture is as follows.

(1) a. Is it the case that p? Yes! *asserts p*, $+p$
b. Is it the case that p? No! *rejects p*, $-p$
c. Is it the case that not p? Yes! *asserts not p*, $+\neg p$

Taking *no* to indicate rejection, however, leads to a different problem, first raised by Imogen Dickie (2010).

Several inferentialists take legitimate inference to preserve evidence. Dag Prawitz holds that a deductive inference is legitimate only if

> a subject who makes the inference and has evidence for its premisses thereby gets evidence for the conclusion (Prawitz 2015, 73)

and since

> assertions are evaluated among other things with respect to the *grounds* or *evidence* the speakers have for making them, we may also say that the aim of . . . inferences is to make assertions *justified*. (Prawitz 2015, 71)

Michael Dummett (1991, 176) similarly claims that "deductive argument . . . preserves some property of statements that renders an assertion of them correct". On his view, inference proceeds by "rearranging" what justifies asserting the premises to obtain justification to assert the conclusion. He conceives of this as particular arrangements of observations and mathematical facts, corresponding to canonical processes of verification (1978, 308; 1991, 176, 211, 317–18; see also Dickie 2010, 164). However, such a verificationist conception of evidence and justification is not part and parcel of the view that inference preserves evidence.

According Dummett's and Prawitz's views, inference is tightly connected with the justification of assertions. One must be able to provide evidence to justify one's assertions and such evidence is preserved in legitimate inference, which serves to justify further assertions. Crucially, inferring is something one *does* to obtain evidence. One does not simply possess evidence for some claim A in virtue of possessing evidence for simpler claims that happen to entail A. Someone having evidence for some premises $A_1, ..., A_n$ that entail (possibly by a highly involved inference) some conclusion does not thereby have evidence to assert that conclusion. To have such evidence, they must have *obtained* it, for instance by carrying out the inference themselves or by knowing that there is such an inference, for example from reliable testimony. In the latter case, one has evidence for the premises and for the claim that *if* $A_1, ..., A_n$*, then* A. By applying *modus ponens*, one then obtains evidence to assert A.

The notion of evidence featuring in this explanation of inference is in line with the standard idea in epistemology that evidence matters for the justification of beliefs (and hence assertions). Besides this, the idea that inference preserves evidence is compatible with several understandings of evidence. For instance, evidence could be understood in a broadly evidentialist way (Conee and Feldman 2004), so that *only* evidence can justify assertions. Or it could be identified with rational credence or subjective probability, so that an inference preserves evidence if the subjective probability one assigns to all premises (jointly) is as least as high as the subjective probability one assigns to the conclusion (Schulz 2010). Yet another option would be to treat *evidence* as identical with *knowledge* and endorse a norm of assertion to establish the connection between epistemology and assertoric practice (Williamson 2000).

Now, Dickie argues that bilateralists cannot hold that inference preserves evidence unless they give up on their claim to be able to vindicate classical logic. According to the inference-preserves-evidence view, the evidence to justify an assertion is *specific*: it is evidence for the particular sentence being asserted. Dickie observes that inspection of negative answers to polar questions reveals that rejections are *unspecific* with respect to the evidence that justifies their performance. A sentence may be properly rejected on the basis of one being justified to assert its negation, but need not be, as witnessed by the following examples. The first two are due to Dickie (2010), the third is adapted from Incurvati and Schlöder 2017.

(2) Is it the case that Homer wrote the *Iliad*? No! Homer did not exist.

(3) Is it the case that Socrates was a unicorn? No! There is no such thing as the property of being a unicorn.

(4) Is it the case that X will win the election? No! X or Y will win.

These rejections are *weak*: they can be correctly performed by someone not having evidence for the negation of the sentence being rejected. For instance, in (4) the speaker is not rejecting *X will win the election* on the basis of evidence that X will not win, but on the basis of evidence that X or Y will win. Similar considerations apply to the other examples. Thus, it is indeterminate whether a sentence is rejected on the basis of evidence for its negation or on some other basis, such as evidence for another sentence. Dickie concludes that their lack of association with any particular kind of evidence makes rejections unsuitable to serve as premises and conclusions in an evidence-preserving proof theory. Assertions are associated with specific evidence that justifies their performance and that is preserved in legitimate inference. In contrast, the association of rejections with evidence is unspecific, so it is indeterminate what would be preserved in an inference featuring a rejection among its premises.

Bilateralists might insist that they only intended to talk about *strong* rejections—rejections that are made on the basis of evidence for a sentence's negation. This, Dickie continues, would not get bilateralists out of their predicament. For if the force marker '$-$' stands for strong rejections only, the Smileian *reductio* principle (SR_1) does not preserve evidence and is therefore invalid. For instance, it would be absurd for someone having evidence that Homer does not exist to assert that *Homer wrote the Iliad*. By Smileian *reductio*, they might then reject *Homer wrote the Iliad*. If this rejection is strong and inference preserves evidence, they could by this inference obtain evidence for *Homer did not write the Iliad*. But from evidence for *Homer did not exist* one cannot inferentially obtain evidence for *Homer did not write the Iliad*.[7]

The problem is that, like rejections, inferences towards absurdity are unspecific with respect to evidence. It may be absurd for a speaker to assert A because they have evidence for its negation. In this case, they would not be mistaken to strongly reject A. But it may also be absurd for a speaker to assert A while it would be a mistake for them to strongly reject A. In such cases, they may only weakly reject A. Hence, (SR_1) is not valid for strong rejections.

Bilateralists are confronted with a dilemma. If Smileian *reductio* is valid for their rejection sign, then the notion of rejection encompasses weak rejections, which cannot serve as premises and conclusions in an evidence-preserving proof theory. If Smileian *reductio* is not valid for their rejection sign, then *reductio* is not valid either.[8] Either way, bilateralists cannot have a classical evidence-preserving proof theory.

[7] One may think that one *can* obtain justification for a negation of this sentence, namely for *It is not the case that Homer wrote the Iliad*. To say that the rejection of *Homer wrote the Iliad* is made on the basis of justification for its external/wide-scope negation, one would need to endorse a quantificational approach to proper names such as *Homer*. This is, for good reasons, a fringe view. We agree with the consensus that proper names trigger presuppositions in, essentially, the Frege–Strawson sense. An alternative approach, compatible with the presupposition view, is to say that the negation in *It is not the case that Homer wrote the Iliad* is metalinguistic. But metalinguistic negation does not appear to behave classically: for instance, one can metalinguistically reject classical tautologies such as *Homer wrote the Iliad or Homer did not write the Iliad*. So Dickie's point against the bilateralist's defence of classical logic stands.

[8] This can be seen as follows. Suppose that the assertion of A leads to absurdity. By *reductio*, we can infer the assertion of $\neg A$. If negative assertion implies rejection—as it should—then we can obtain the rejection of A. That is, (SR_1) is valid. Thus, if (SR_1) is invalid, then so is *reductio*.

In the next section, we argue that bilateralists should tackle the dilemma by having *two* rejection signs. For one of them, Smileian *reductio* is valid. This means that the sign stands for rejections that can be weak. Nonetheless, weak rejections can be accommodated within proof theory by taking inference to preserve commitment, instead of evidence. For the other rejection sign, Smileian *reductio* is not valid. However, this is the case only for inferences to absurdity that fail to preserve evidence. By restricting attention to evidence-preserving inferences to absurdity, bilateralists can have a classical evidence-preserving proof theory as a fragment of their extended proof theory that includes weak rejections and preserves commitment. This extended proof theory itself is almost classical: although *reductio* fails because of the presence of weak rejections, the proof theory nevertheless validates all classically valid arguments.

3 Negation and Evidence

Dickie (2010) observed that rejections *tout court* (weak or strong) are unspecific with respect to evidence: it is indeterminate which assertions are justified when it is correct for one to reject a sentence. For this reason, rejections cannot serve as premises and conclusions in evidence-preserving inferences. In Incurvati and Schlöder 2017, we reply that there is nevertheless a good sense of inference to which rejections can contribute: the preservation of *commitment*. Dickie situates her discussion in the context of a literature focusing on the evidence for assertions, but in Incurvati and Schlöder 2017 we stress the alternative option of focusing on the commitments undertaken by assertions (see, for instance, Brandom 1994). One's commitments are part of, in Lewis's (1979) terminology, the conversational scoreboard. These commitments entail permissions and obligations by the rules of the language game. If, say, these rules prohibit self-contradiction, then someone who committed to A and then also commits to $\neg A$ is obliged to retract one of these commitments once this is pointed out to them. We go on to define commitment-preserving inference as follows.

> Given that a speaker has undertaken certain commitments ... [inference]

> rules tell us what further commitments that speaker is bound by. (Incurvati and Schlöder 2017, 8)

For example, a bilateral rule for conjunction states that $+(A \wedge B)$ entails $+A$. This rule preserves commitment, since whenever a speaker is committed to $A \wedge B$, they are also committed to A: if A is up for discussion, they must grant A or admit to a mistake and retract their commitment to $A \wedge B$. Speakers need not be aware of all their commitments. Indeed, it is cognitively implausible that they can be, since infinitely many sentences follow from any given A. But if one of their commitments is pointed out to them, speakers are obliged to grant the conclusion or admit to a mistake. These obligations are what is preserved in an inference that preserves commitment (see also Dutilh Novaes 2015).

Why, however, should there be a difference between inferences that preserve justification and those that preserve commitment? Put differently, what could be the function of commitment, if one can be committed without having justification? We find an answer by considering an example by Gilbert Harman (1986).

(5) a. If needles are being stuck into a wax doll shaped like me, I am in intense pain.
b. Needles are being stuck into a wax doll shaped like me.

---

c. I am in intense pain.

One cannot, Harman observes, reason oneself into feeling pain, no matter how convinced one is of folk voodoo. This inference, in our terminology, does not preserve evidence, since even if someone could have justification to assert both premises, they would not thereby have justification to assert the conclusion. Nevertheless the inference is valid in that it preserves commitment. That this is so fulfills an important social function. A persistent interlocutor may get someone who believes in folk voodoo to assert both premises and then point out to them that they hence *ought* to concede the conclusion that they feel pain when in fact they do not. One salient purpose is to prompt a re-evaluation of the premises; another to ridicule the speaker in front of overhearers.

The view that legitimate inference preserves commitment makes room for rejections that may be weak. When one rejects A, one is making explicit that one expresses one's refraining from committing to A—one abnegates any obligation to grant A.[9] As Dickie noted, Smileian *reductio* is valid for rejections that can be weak. If from hypothetically committing to A (that is, $+A$), absurdity follows (that is, $\bot$), one can infer that one need not grant A. Conversely, if from the hypothesis that one need not grant A it follows that $\bot$, one can infer that one is implicitly committed to A. In addition, it is absurd to commit to A and refrain from doing so. Thus, the coordination principles preserve commitment when taken to be about rejections *tout court* (Incurvati and Schlöder 2017, 9). Hence, using $\ominus$ as a sign for rejections *tout court*, the following rules preserve commitment.

$$(\text{Rejection})\ \dfrac{+A \qquad \ominus A}{\bot} \qquad (\text{SR}_1)\ \dfrac{\begin{array}{c}[+A]\\ \vdots\\ \bot\end{array}}{\ominus A} \qquad (\text{SR}_2)\ \dfrac{\begin{array}{c}[\ominus A]\\ \vdots\\ \bot\end{array}}{+A}$$

However, rejections *tout court* do not validate all bilateral negation rules. In particular, it does not follow from the fact that one refrains from committing to A that one is committed to *not* A. And it does not follow from the fact that one refrains from committing to *not* A that one is committed to A. Thus, $(+\neg\text{I.})$ and $(-\neg\text{E.})$ do not preserve commitment for rejections *tout court*. Nonetheless, like the other bilateral negation rules, they do preserve commitment for strong rejections. Hence, reserving the $-$ sign for strong rejections, the following rules preserve commitment.

$$(+\neg\text{I.})\ \dfrac{-A}{+\neg A} \qquad (+\neg\text{E.})\ \dfrac{+\neg A}{-A} \qquad (-\neg\text{I.})\ \dfrac{+A}{-\neg A} \qquad (-\neg\text{E.})\ \dfrac{-\neg A}{+A}$$

Dickie's first observation was that rejections are unspecific with respect to evidence. Understanding inference in terms of commitment preservation allows one to include rejections in one's proof theory despite their unspecificity. However, Dickie's second observation was that inferences towards absurdity are likewise unspecific, which means that Smileian *reductio* for strong rejections does not preserve evidence. Given that we are no longer understanding inference in terms of evidence preservation, it does not

[9] Note the difference, familiar from the expressivist literature in metaethics, between expressing and reporting an attitude. By weakly rejecting, one is *expressing* one's refraining from committing. This is different from *reporting* that one refrains from committing. The utterance *I refrain from committing to A* serves to perform not a weak rejection, but an assertion reporting an attitude.

follow that Smileian *reductio* for strong rejections is invalid. However, the same example we used above to show that Smileian *reductio* for strong rejections does not preserve justification shows that Smileian *reductio* for strong rejections does not preserve commitment either: it is absurd for someone committed to believing *Homer did not exist* to commit to believing *Homer wrote the Iliad*, but it does not follow that they are committed to believing *Homer did not write the Iliad*. Since the commitment-preserving rules governing negation do not deliver the classical laws of negation without the coordination principles, it appears that the bilateralist cannot have a classical *commitment-preserving* proof theory either.

Not all is lost, however. Even if Smileian *reductio* is invalid for strong rejections when unspecific inferences to absurdity are countenanced, it may still be valid when such unspecific inferences are excluded. It turns out that adding appropriately restricted versions of Smileian *reductio* for strong rejections to the commitment-preserving proof theory suffices to validate all classically valid arguments. Moreover, the evidence-preserving fragment of the resulting proof theory obeys classical logic.

When dealing with the unspecificity of rejection, we isolated the evidentially specific instances—the strong rejections—and noted that the bilateral negation rules preserve commitment, despite the fact that they fail to do so for rejection *tout court*. Similarly, we can isolate the evidentially specific inferences towards absurdity. For those inferences, Smileian *reductio* for strong rejection is valid in that it preserves commitment and indeed evidence, despite the fact that it is not valid for inferences towards absurdity *tout court*. The inferences towards absurdity that are evidentially specific include at least the inferences that preserve evidence, as an inference that proceeds from evidentially specific premises and preserves evidence cannot reach an evidentially unspecific conclusion. Thus, the following *Smileian reductio** rules, which together with the (Strong Rejection) rule form the *coordination principles**, preserve both commitment and evidence.

$$(\text{S. Rej.})\ \frac{+A \quad -A}{\bot} \qquad (\text{SR}_1{*})\ \frac{\begin{array}{c}[+A]\\ \vdots \\ \bot\end{array}}{-A}\ \begin{array}{l}\text{if the inference to } \bot\\ \text{preserves evidence}\end{array} \qquad (\text{SR}_2{*})\ \frac{\begin{array}{c}[-A]\\ \vdots \\ \bot\end{array}}{+A}\ \begin{array}{l}\text{if the inference to } \bot\\ \text{preserves evidence}\end{array}$$

These rules are not subject to Dickie's counterexamples to Smileian *reductio* for strong rejections. For example, the inference towards absurdity from *Homer did not exist* and the assumption *Homer wrote the Iliad* does not preserve evidence. Homer's existence is a precondition for intelligible talk of evidence for him having written the Iliad. Thus if *Homer did not exist* is a premise, *Homer wrote the Iliad* cannot occur in an evidence-preserving argument at all. This is not to say that it would be incorrect to infer absurdity from *Homer did not exist* and the assumption *Homer wrote the Iliad*. This inference is valid in that it preserves commitment. For someone who commits to *Homer wrote the Iliad* is committed to Homer's existence, which is incompatible with commitment to *Homer did not exist*. But it does not preserve evidence, so it excluded from Smileian *reductio** and one cannot infer the strong rejection of *Homer wrote the Iliad* on its basis.

The Smileian *reductio** rules are formulated by restricting their inferences to absurdity to those that preserve evidence. The question is which inferences preserve evidence and how they can be characterized in a way that can be formally stated in an inference rule. Within the confines of the language of propositional logic, the answer is simple. Failures of evidence preservation may only arise because of the presence of weak rejections. Thus if an inference involves no weakly rejected premises, it trivially preserves evidence. We can therefore phrase the Smileian *reductio** rules as follows.

$$(\mathrm{SR}_1{*})\ \dfrac{\begin{matrix}[+A]\\ \vdots\\ \bot\end{matrix}}{-A}\ \begin{matrix}\text{if no premises signed with }\ominus\\ \text{were used to derive }\bot\end{matrix}\qquad (\mathrm{SR}_2{*})\ \dfrac{\begin{matrix}[-A]\\ \vdots\\ \bot\end{matrix}}{+A}\ \begin{matrix}\text{if no premises signed with }\ominus\\ \text{were used to derive }\bot\end{matrix}$$

Indeed, this is what we do in Incurvati and Schlöder 2017 when we claim that the following is the valid version of *reductio* in bilateral logic (although we do not frame the argument as being about the preservation of evidence).

$$(\text{Bilateral } \textit{Reductio})\ \dfrac{\begin{matrix}[+A]\\ \vdots\\ \bot\end{matrix}}{+\neg A}\ \text{if no premises signed with }\ominus\text{ were used to derive }\bot$$

Let *weak bilateral logic* (WBL) be the natural deduction calculus consisting of the bilateral negation rules, the coordination principles, the coordination principles* and the following rules for conjunction.

$$(+\wedge\text{I.})\ \frac{+A \qquad +B}{+A \wedge B} \qquad (+\wedge\text{E.}_1)\ \frac{+A \wedge B}{+A} \qquad (+\wedge\text{E.}_2)\ \frac{+A \wedge B}{+B}$$

We write $\vdash^{\mathsf{WBL}}$ for the inference relation defined by this calculus and $\models^{\mathsf{CPL}}$ for the consequence relation of classical propositional logic.

Theorem 3.1 (Incurvati and Schlöder 2017). $\Gamma \models^{\mathsf{CPL}} A$ *iff* $\{+X \mid X \in \Gamma\} \vdash^{\mathsf{WBL}} +A$.

Thus, the valid arguments in the logic of assertion are exactly the classically valid arguments. This means in particular that formulating versions of Smileian *reductio* that are valid for strong rejection enables the bilateralist to vindicate all the classically valid arguments. However, as the entirety of WBL has expressive power beyond classical logic due to the inclusion of weak rejections, it does not follow that all classical *meta-rules* are valid. For example, it is not in general the case that if $\Gamma, +A \vdash^{\mathsf{WBL}} \bot$, then $\Gamma \vdash^{\mathsf{WBL}} +\neg A$ if Γ contains weakly rejected premises. Thus the meta-rule of *reductio* fails in WBL. As shown by the derivability of Bilateral *Reductio* in WBL, however, this failure can *only* arise in the presence of unspecific inferences to absurdity: the evidence-preserving fragment of WBL is fully classical.

Once we extend the object language beyond propositional logic, it becomes more difficult to determine which inferences preserve evidence. Moritz Schulz (2010) gives an example involving epistemic *must*, formalized as $\Box$. A plausible inference involving *must* is *epistemic strengthening*: from A, infer *it must be that* A. The intuitions underwriting this inference appear to be solid, but extending classical propositional logic so that $A \models \Box A$ immediately leads to disaster (Yalcin 2007). For if one lets $\Diamond$ abbreviate $\neg\Box\neg$ then a *reductio* argument using epistemic strengthening shows $\Diamond A \models A$ for any A, trivializing the modal.

Schulz argues that one should have never considered epistemic strengthening to be valid in the first place. He notes that there are situations in which one has strong evidence for some claim A, but less or no evidence for *it must be that* A. Suppose, for instance, that one sees that the lights are on. Then one seems to have strong evidence for *They are home* but less or no evidence for *They must be home* (see also Bledin and Lando 2018 for analogous examples). Schulz concludes that epistemic strengthening does not preserve evidence and should be rejected. The bilateralist response to Dickie extends to a response to Schulz: although epistemic strengthening does not preserve evidence (so it cannot occur in Smileian *reductio** arguments), it preserves commitment. As we note in a later paper (Incurvati and Schlöder 2022), the approach of adding epistemic strengthening to bilateral logic, but excluding it from Smileian *reductio**, accounts for the available data about epistemic modality while also preserving the intuitive appeal of epistemic strengthening and the classical laws of negation. Notably, sentences like *it is raining and it might not be raining* are provably contradictory (as they should be), but *it might be A* does not in general entail *A*.

4 Truth and Evidence

Price gives a functionalist explanation of the negation operator. The *raison d'être* of negation is to fulfill a particular expressive need, namely to register perceived incompatibilities by allowing us to perform rejections. Deflationists about truth give a similarly functionalist explanation of the truth predicate. The *raison d'être* of the truth predicate is to fulfill a particular expressive need, namely to indirectly and compendiously endorse by performing assertions such as the following.

(6) The sentence written on Thomas's whiteboard is true. (indirect endorsement)

(7) Everything Thomas says is true. (compendious endorsement)

With a truth predicate, we can perform assertions expressing indirect and compendious endorsements. The deflationist goes further and claims that our languages contain a truth predicate *so that* we can make such endorsements. Allowing us to make indirect and compendious endorsements is the function of the truth predicate and its meaning is to be explained—solely and exactly—by appealing to this function.

The deflationist story about the truth predicate is well known. What is perhaps less well known is that, just like the functionalist story about negation, it is subject to the Frege-Geach problem: the truth predicate can appear in contexts, such as conditional antecedents, where it cannot be plausibly taken to express endorsement. For instance, one may utter the following sentence if one knows that Thomas's whiteboard is exclusively used for putative counterexamples to Goldbach's conjecture but it is unlikely that Thomas has provided such a counterexample.

(8) If the sentence written on Thomas's whiteboard is true,
there is a counterexample to Goldbach's conjecture.

Earlier, we saw that functionalists about negation can solve their Frege-Geach problem by embedding their account within the bilateral framework. Although negation does not express incompatibility, its meaning is inferentially explained in terms of rejection, which expresses incompatibility. Deflationists, we argue, should follow suit. Although the truth predicate does not indicate endorsement, its meaning is inferentially explained in terms of *assertion*, which expresses endorsement. Accordingly, we take the meaning of the truth predicate to be given by *inference rules* encoding its relation to assertion. These *asserted truth rules* state that from an assertion of $\ulcorner A \urcorner$ *is true* one can infer an assertion of A and *vice versa*.

$$(\text{+T-IN})\ \frac{+A}{+T\ulcorner A\urcorner} \qquad (\text{+T-OUT})\ \frac{+T\ulcorner A\urcorner}{+A}$$

The resulting view is a form of *inferential deflationism*, the view that the meaning of the truth predicate is given by rules allowing one to pass from A to $\ulcorner A \urcorner$ *is true* and *vice versa*.

Our inferential deflationism addresses the Frege-Geach problem for truth analogously to how bilateralism addressed the Frege-Geach problem for the functionalist account of negation. In uttering (8), one is indeed not endorsing the sentence on Thomas's whiteboard. But the utterance does commit one to there being a counterexample to Goldbach's conjecture should one endorse the sentence on Thomas's whiteboard by uttering *the sentence on Thomas's whiteboard is true*. The asserted truth rules ensure this by validating the appropriate inferences. Similarly to the case of negation, we can accommodate the Frege-Geach point that in certain contexts the truth predicate does not express endorsement. Nonetheless, the meaning of the truth predicate is explained (via the asserted truth rules) in terms of its inferential relation to assertion, which expresses endorsement.

But now paradox looms: if we add the asserted truth rules to weak bilateral logic, we can use the Liar sentence L to derive a paradox. We first use the asserted truth rules to show that the Liar is interderivable with its own negation. To wit, using (+T-IN) we can show that $+L \vdash^{\mathsf{WBL}} +\neg L$.

$$\dfrac{\dfrac{+L}{+T\ulcorner L\urcorner}\,(+\text{T-IN}) \qquad +L \leftrightarrow \neg T\ulcorner L\urcorner}{+\neg L}\,(\text{by contraposition})$$

And using (+T-OUT) we can show that $+\neg L \vdash^{\mathsf{WBL}} +L$.

$$\dfrac{\dfrac{+\neg L \qquad +L \leftrightarrow \neg T\ulcorner L\urcorner}{+T\ulcorner L\urcorner}\,(\text{by contraposition})}{+L}\,(+\text{T-OUT})$$

Then, since $+L \vdash^{\mathsf{WBL}} +\neg L$ and negation registers incompatibility, it follows that $+L \vdash^{\mathsf{WBL}} \bot$. By Smileian *reductio**, it follows that $\vdash^{\mathsf{WBL}} -L$. This entails $\vdash^{\mathsf{WBL}} +\neg L$ by the bilateral rules for negation. But since $+\neg L \vdash^{\mathsf{WBL}} +L$, we can conclude that $\vdash^{\mathsf{WBL}} \bot$.

The standard reaction is to blame the asserted truth rules, which are used to establish the interderivability of the Liar and its negation. The asserted truth rules, however, are central to our inferential deflationist explanation of the meaning of the truth predicate and to our solution to the Frege-Geach problem. Hence, these rules should not be lightly given up. The standard reaction is however correct that the interderivability of the Liar and its negation deserves closer scrutiny.

On the inference-preserves-evidence view, the interderivability of the Liar and its negation shows that evidence for L entails evidence for $\neg L$, and evidence for $\neg L$ entails evidence for L. If it is possible at all to have evidence for L or evidence for $\neg L$, this means that it is possible for the same evidence to support both L and $\neg L$. This clashes with the idea that negation registers incompatibility.

One option is to say that there cannot be evidence for L or $\neg L$. This is a version of the paracomplete solution to the semantic paradoxes (Field 2008), which entails rejecting the law of excluded middle. Another option is to abandon the idea that negation registers incompatibility and say that there can be evidence for a sentence and its negation. This is a version of the paraconsistent solution (Priest 1979; Beall 2009), which entails rejecting the law of non-contradiction.

There is a third option. Within our framework, inference need not preserve evidence, but only commitment. The framework therefore opens up the possibility of holding on to the interderivability of the Liar and its negation whilst retaining the laws of contradiction and excluded middle. We can take the asserted truth rules to be valid in that they preserve commitment, but insist that they do not preserve evidence. As a result, the inferences from L to $\neg L$ and from $\neg L$ to L, albeit valid, do not preserve evidence either. This allows us to give a new diagnosis of the Liar paradox: what goes wrong in its derivation is that it applies Smileian *reductio** to inferences that only preserve commitment, whereas Smileian *reductio** can only be legitimately applied to inferences that preserve evidence.

In support of our diagnosis, we now present an independent argument to the effect that the asserted truth rules do not preserve evidence. Consider again (6).

(6) The sentence written on Thomas's whiteboard is true.

Suppose the sentence written on Thomas's whiteboard is *it is raining*. If (+T-OUT) preserves evidence, then from one having evidence for *the sentence on Thomas's whiteboard is true* it follows that one can inferentially obtain evidence for the sentence on Thomas's whiteboard, that is evidence for *it is raining*. But this does *not* follow. For suppose that someone has evidence for *the sentence on Thomas's whiteboard is true* because they know that Thomas (for whatever reason) writes a true sentence on his whiteboard every day. They do not know which truth is written on the board, but this does not change the fact that they have evidence for *the sentence on Thomas's whiteboard is true*: they could vindicate their assertion of this sentence by appealing to their knowledge. But they may be unaware of the weather and unaware that *it is raining* is the sentence on Thomas's whiteboard. So they could not inferentially obtain justification for *it is raining*.[10] So

[10] On a Williamsonian way of spelling out the relationship between evidence and inference (Williamson 2000), one could say that they *do* have evidence for *it is raining*, but do not know that they do—*because* they have evidence for *the sentence written on Thomas's whiteboard is true* and (unbeknownst to them) the sentence written on Thomas's whiteboard is *it is raining*. This is not how we understand inference. One does not just *have* evidence because some inference is valid, but *obtains* evidence *by* an inference. In the case under discussion, the speaker cannot use the inference to obtain evidence, since they cannot even phrase this inference, as they do not know what goes in the conclusion. To stress, our view is compatible with the view that evidence is knowledge and other broadly externalist conceptions.

(+T-OUT) does not preserve evidence.[11]

We can make the argument slightly more formal by letting $\ulcorner W \urcorner$ be a name for the sentence W on Thomas's whiteboard. Someone familiar with Thomas's habit of writing true sentences on his whiteboard has evidence for $\ulcorner W \urcorner$ *is true* and may assert this, that is $+T\ulcorner W \urcorner$. However, they may not have evidence for W. This is because someone who asserts that $\ulcorner W \urcorner$ *is true* need not know what sentence is denoted by $\ulcorner W \urcorner$ and hence may not be able to justify asserting the *particular* sentence denoted by $\ulcorner W \urcorner$. But this particular sentence is just W, so they may not be able to justify asserting W. But being able to justify asserting $+T\ulcorner W \urcorner$ but not being able to justify asserting $+W$ means that inferring $+W$ from $+T\ulcorner W \urcorner$ does not preserve evidence.

The converse argument shows that (+T-IN) does not preserve evidence either. In brief, suppose that one has evidence for some sentence W and it just so happens that W is also the sentence on Thomas's whiteboard, but this is unknown (and, in this case, one has no prior evidence regarding Thomas's habits). In this situation, one has evidence for the sentence that is denoted by *the sentence on Thomas's whiteboard*, so one has evidence for the sentence on Thomas's whiteboard. But one cannot inferentially obtain evidence for *the sentence on Thomas's whiteboard is true*: one cannot justify its assertion. Thus (+T-IN) does not preserve evidence.

Although the asserted truth rules do not preserve evidence, they preserve commitment. Consider (+T-OUT). If someone asserts *the sentence written on Thomas's whiteboard is true* and it turns out that the sentence on Thomas's whiteboard is W, they are committed to W: they are conversationally obliged to concede W once all the relevant facts are known or admit to a mistake. The converse shows that (+T-IN) preserves commitment.

We take part with the Williamsonian idea when it comes to the anti-luminous properties governing the relation between evidence and inference.

[11] A similar argument related to (+T-OUT) has recently been discussed by Daniel Drucker (2020). He argues that it is sometimes rational to believe of some proposition H that $\ulcorner H \urcorner$ *is true*, but irrational to believe H. He develops a theory of belief that predicts these judgements of rationality and irrationality, but does not explain *why* the application of (+T-OUT) here is a mistake. Drucker seems forced to reject (+T-OUT), yet he appears to endorse it. Our explanation—that (+T-OUT) preserves commitment, but not evidence—may close this explanatory gap and be aligned with Drucker's theory. To wit: if it is rational to believe p and p *entails* q, it follows that it is rational to believe q only if the inference of p from q preserves evidence, which (+T-OUT) does not. We will not pursue this point further, as Drucker's concerns (on the differences between rational belief and rational desire) are tangential to ours.

If someone asserts that W, they are conversationally obliged to agree that *the sentence written on Thomas's whiteboard is true* once it is pointed out to them that W is written on Thomas's whiteboard. The reason for the difference between commitment and evidence here is that evidence is dependent on the epistemic state of individual speakers who may be ignorant of certain facts (such as the denotations of names), whereas commitments are externally imposed onto them, taking such facts into account.

An objection to our argument begins by pointing out that the case of the Liar is different from our example involving Thomas's whiteboard, which involves uncertainty about what sentence is denoted by a name. There is no such uncertainty about $\ulcorner L \urcorner$ denoting L. Thus, the objection goes, might it be that the truth rules preserve justification in the Liar paradox, even if they do not preserve justification in all their possible uses?

We have presented the failure of evidence preservation as a diagnosis of the Liar paradox that is at least on all fours with the paracompletist and paraconsistentist diagnoses. Paraconsistentists cite the Liar paradox to support their claim that there are glutty sentences; paracompletists cite it to support their claim that there are gappy sentences. We cite the Liar paradox to support the claim that the truth rules do not preserve evidence and be on a par with these approaches. The paradox shows that these rules cannot be used under *reductio* as much as it shows that they are gluts or gaps. In addition, we observe that the example involving Thomas's whiteboard shows that there are reasons independent of the semantic paradoxes to hold that the truth rules do not preserve evidence.

Moreover, it is worth stressing that, like the property of truth preservation in classical logic, evidence preservation and commitment preservation are, strictly speaking, properties of rules, not of particular inferences. Thus, the fact that there is one case where one may have evidence for the premise of a truth rule without having evidence for its conclusion suffices to establish that the truth rules are not evidence preserving. That there are instances of the truth rules, such as the inference from *5+7=12* to *'5+7=12' is true*, in which one has evidence for the conclusion whenever one has evidence for the premise is neither here nor there. Having said this, it is of course possible to consider whether the truth rules preserve evidence when restricted to applications of the truth predicate to direct quotations. This would however be inadvisable for deflationists. For it would amount to completely giving up on their usual story about the function of the truth predicate as a device for indirect and compendious endorsement. If, as deflationists claim, the truth rules wholly determine the meaning of the truth predicate and, as the putative objection goes, the truth rules only apply to direct quotations, then the meaning of *the sentence on Thomas's whiteboard is true* is not explained.[12]

Thus, our diagnosis is that the Liar Paradox applies the asserted truth rules as if they preserved evidence, but these rules only preserve commitment. This diagnosis leads to a cure. Since the asserted truth rules preserve commitment, we may add them to weak bilateral logic, as desired. Nonetheless, since these rules do not preserve evidence, we must exclude their application from Smileian *reductio**. We show below that doing so blocks the derivation of the paradox.

Formally, we extend the language of WBL with a truth predicate and add (+T-IN) and (+T-OUT) to the system. Furthermore, we must add rules ensuring that truth is also disquotational under strong rejection. Kevin Scharp (2013, 63) observed that the truth predicate is not only needed for indirect and compendious endorsement, but also for expressing indirect and compendious opposition. In the bilateral framework, one can give due to this observation by appropriately relating truth to strong rejection, the speech act expressing strong opposition, by adopting the following *strongly rejected truth rules*.

$$(-\text{T-IN})\ \frac{-A}{-T\ulcorner A\urcorner} \qquad (-\text{T-OUT})\ \frac{-T\ulcorner A\urcorner}{-A}$$

These rules do not preserve evidence, for the same reason that their asserted analogues do not. They do however preserve commitment, again for the same reason that their

[12] One may restrict the truth rules to direct quotations and, in order to explain indirect endorsement, attempt to validate inferences of the form *The sentence on Thomas's whiteboard is true; The sentence on Thomas's whiteboard is 'It is raining'; therefore it is raining*. This cannot work, however, since with the truth rules restricted to direct quotation only, the meaning of the first premise in such arguments is unexplained.

asserted analogues do. The *truth rules*—that is, the asserted truth rules and the strongly rejected truth rules—jointly entail that truth is bivalent, that is that $+T\ulcorner\neg A\urcorner$ is derivable from $+\neg T\ulcorner A\urcorner$.

$$\dfrac{\dfrac{\dfrac{\dfrac{+\neg T\ulcorner A\urcorner}{-T\ulcorner A\urcorner}\,(+\neg\text{E.})}{-A}\,(-\text{T-OUT})}{+\neg A}\,(+\neg\text{I.})}{+T\ulcorner\neg A\urcorner}\,(+\text{T-IN})$$

Scharp's example of opposition expressed by using the truth predicate is that in uttering *the continuum hypothesis is not true*, one can oppose the continuum hypothesis. As seen in the first three steps of the above derivation, it is indeed the case that if one asserts *the continuum hypothesis is not true* it follows that one strongly rejects the continuum hypothesis. In our bilateral account, one can achieve the same result by strongly rejecting *the continuum hypothesis is true*.

Continuing Scharp's line of reasoning, we may also consider weak opposition as expressed by weak rejections and simply call opposition what is expressed by rejections *tout court*. Then, the truth predicate is also needed for indirect or compendious opposition. It is not necessary to add further truth rules to account for this, as the following are *derivable* from the asserted truth rules using Smileian *reductio*.

$$(\ominus\text{T-IN})\ \dfrac{\ominus A}{\ominus T\ulcorner A\urcorner} \qquad (\ominus\text{T-OUT})\ \dfrac{\ominus T\ulcorner A\urcorner}{\ominus A}$$

If we extend the language of weak bilateral logic to include a truth predicate governed by the truth rules, we must take care to phrase the strong versions of Smileian *reductio* correctly. That is, we must ensure that they only apply to subderivations that preserve evidence:

$$(\text{SR}_1{*})\ \dfrac{\begin{matrix}[+A]\\ \vdots\\ \bot\end{matrix}}{-A}\ \begin{matrix}\text{if the inference to } \bot \text{ uses no premis-}\\ \text{es signed with } \ominus \text{ and no truth rules}\end{matrix} \qquad (\text{SR}_2{*})\ \dfrac{\begin{matrix}[-A]\\ \vdots\\ \bot\end{matrix}}{+A}\ \begin{matrix}\text{if the inference to } \bot \text{ uses no premis-}\\ \text{es signed with } \ominus \text{ and no truth rules}\end{matrix}$$

Call the resulting system $\mathsf{WBL_T}$. Its treatment of the Liar Paradox is as follows. From the fact that $+L \vdash^{\mathsf{WBL_T}} +\neg L$ it follows that $+L \vdash^{\mathsf{WBL_T}} \bot$ and thus by Smileian *reductio* that $\vdash^{\mathsf{WBL_T}} \ominus L$. Thus the Liar sentence ought to be rejected. Similarly, it follows from the fact that $+\neg L \vdash^{\mathsf{WBL_T}} +L$ that $+\neg L \vdash^{\mathsf{WBL_T}} \bot$ and so that $\vdash^{\mathsf{WBL_T}} \ominus\neg L$. Thus the Liar's negation ought to be rejected too. But $\ominus L$ and $\ominus\neg L$ are jointly consistent. And since the inferences towards $\ominus L$ and $\ominus\neg L$ apply a rule that is not evidence preserving, we cannot infer $+\neg L$ or $+L$ by Smileian *reductio**. In the Appendix, we prove that Liar sentences are consistent in $\mathsf{WBL_T}$.[13]

Terence Parsons (1984) and Mark Richard (2008) too argued that Liar sentences and their negations ought to be rejected. But their notions of rejection are stronger than the notion of rejection *tout court* we are adopting from our earlier work (Incurvati and Schlöder 2017). Notably, rejecting some sentence and later asserting it would involve a revision of one's commitments according to Parsons or Richard. This is not so on our approach. Rejection *tout court* makes it explicit that one refrains from committing. But one can come to stop refraining from committing simply by undertaking a new commitment, without revising one's extant commitments.

One distinctive feature of our approach is that it reconciles the truth rules with the classical laws of negation. The following is an immediate consequence of the result that the valid inferences involving only asserted premises and conclusions in WBL are exactly the classically valid inferences (Theorem 3.1). Given a map σ from propositional atoms to the sentences of $\mathsf{WBL_T}$ and a propositional logic formula A, let $\sigma[A]$ denote the $\mathsf{WBL_T}$ sentence obtained by uniformly substituting every atom p in A with $\sigma(p)$.

Theorem 4.1. *Let σ map propositional atoms to sentences of* $\mathsf{WBL_T}$. *Then* $\Gamma \models^{CPL} A$ *iff* $\{+\sigma[X] \mid X \in \Gamma\} \vdash^{\mathsf{WBL_T}} +\sigma[A]$.

That is, the $\mathsf{WBL_T}$ logic of assertion validates all substitution instances of classically valid arguments. In particular, $\mathsf{WBL_T}$ proves all instances of the laws of excluded middle and non-contradiction. It does not, however, validate all the classical *meta-rules*, since for example *reductio* is not generally valid.[14] We discuss the repercussions of this in the following section.

[13] A very similar solution can then be pursued for the Curry paradox. We do so elsewhere (Incurvati and Schlöder forthcoming, ch. 8).

[14] Weir (1996, 14) already recommends that *reductio* be "restricted in a non ad hoc fashion" to free deflationism from the Liar paradox. He considered this to be "a very large enterprise indeed", which may entail that it is "no longer a decidable matter whether a construction is a proof." We have, in a sense, undertaken this enterprise without abandoning decidability. Weir only considered excluding

5 Transparency

A theory of truth should account for the use of the truth predicate in actual inferential practice. But our inferential practice appears to validate the classical meta-inferences of *reductio* and proof by cases, both of which are invalid in $\mathsf{WBL_T}$.

Reductio $\quad +A \wedge \neg T\ulcorner A \urcorner \vdash^{\mathsf{WBL_T}} \bot$, but $\nvdash^{\mathsf{WBL_T}} +\neg(A \wedge \neg T\ulcorner A \urcorner)$

Proof by cases $\quad +A \vdash^{\mathsf{WBL_T}} +\neg A \vee T\ulcorner A \urcorner$ and $+\neg A \vdash^{\mathsf{WBL_T}} \neg A \vee T\ulcorner A \urcorner$, but $+A \vee \neg A \nvdash^{\mathsf{WBL_T}} +\neg A \vee T\ulcorner A \urcorner$

Both counterexamples revolve around the fact that one cannot derive the material conditional $+A \rightarrow T\ulcorner A \urcorner$ despite the validity of $+A \vdash^{\mathsf{WBL_T}} +T\ulcorner A \urcorner$. Hence, the conditional proof rule fails for $\mathsf{WBL_T}$ and it is this failure that is responsible for the failures of the other meta-rules. But this is as it should be: *modus ponens* for the material conditional preserves evidence, but $\vdash^{\mathsf{WBL_T}}$ does not, thus valid inference is not co-extensional with the derivable material conditionals. Hence conditional proof and, with it, the above meta-inferences *should* not be valid. Natural language data does not support the claim that we must accept material biconditionals of the form $+A \leftrightarrow T\ulcorner A \urcorner$. If anything, the data only supports the claim that one must accept the *indicative* conditionals *if A, then* $\ulcorner A \urcorner$ *is true* and *if* $\ulcorner A \urcorner$ *is true, then A*. But the indicative conditional is not the material conditional.[15]

The failure of the meta-rules entails that truth is not fully *transparent*: one cannot intersubstitute $T\ulcorner A \urcorner$ and A in all contexts. Overall, we take this to be a good thing. Denying full transparency plays a crucial role in avoiding Liar paradoxes. Field (2008, §13.3) presents an argument in favor of full transparency, which, on closer inspection is in fact a version of the Frege–Geach problem. The argument is that truth must "be well-behaved ... inside conditionals as in unembedded contexts" (Field 2008, 209-10)

certain classes of sentences from *reductio*, whereas our strategy is to exclude certain *inferences*. On our approach, proof verification remains decidable, as it is decidable whether a subproof contains truth rules. In later work, Weir (2005; 2015) retains *reductio* and deals with the Liar Paradox by making the consequence relation non-transitive. We discuss the relation between our approach and non-transitive approaches to the semantic paradoxes in fn. 16 below.

[15] This does mean that the inferential deflationist should present a semantics for the indicative that supports the assertion of these conditionals. We take up this challenge elsewhere (Incurvati and Schlöder forthcoming, ch. 8).

and that this requires A and $T\ulcorner A\urcorner$ to be intersubstitutable. As shown, we can make sense of truth in conditional antecedents without requiring full transparency. Likewise for versions of the Frege–Geach worry that arise from other kinds of embeddings. We now investigate *how much* of transparency we are giving up.

We were able to recover the valid instances of *reductio* as the restricted rule of Smileian *reductio**. We can do the same for proof by cases. The following rule for disjunction elimination is derivable in weak bilateral logic together with the truth rules.

$$(+\vee\text{E.})\ \dfrac{+A\vee B \qquad \begin{array}{c}[+A]\\ \vdots\\ +C\end{array} \qquad \begin{array}{c}[+B]\\ \vdots\\ +C\end{array}}{+C}\ \begin{array}{l}\text{if both inferences to } +C \text{ use no premises signed with}\\ \ominus \text{ and no truth rules.}\end{array}$$

But does this rule suffice to explain the good inferences in which the truth rules are properly applied under disjunction? For example, from *s or t* and ⌜ *not t*⌝ *is true* it ought to follow that *s*. Intuitively, this inference should go as follows. (The Explosion rule is derivable from Smileian *reductio* using an empty discharge.)

$$\#\ \dfrac{+s\vee t \qquad [+s]^1 \qquad \dfrac{\dfrac{[+t]^2 \qquad \dfrac{\dfrac{+T\ulcorner\neg t\urcorner}{+\neg t}\,(\text{+T-OUT})}{-t}\,(+\neg\text{E.})}{\bot}\,(\text{S. Rejection})}{+s}\,(\text{Explosion})}{+s}\,(+\vee\text{E.})^{1,2}$$

However, this application of disjunction elimination is disallowed by the restrictions on $(+\vee\text{E.})$. Nevertheless, the inference from $+s\vee t$ and $+T\ulcorner\neg t\urcorner$ to $+s$ is valid in $\mathsf{WBL_T}$, as shown by the following derivation.

$$\dfrac{\dfrac{\dfrac{\dfrac{+s\vee t \qquad [+s]^1 \qquad \dfrac{\dfrac{[+t]^2 \qquad \dfrac{[+\neg t]^3}{-t}\,(+\neg\text{E.})}{\bot}\,(\text{S. Rejection})}{+s}\,(\text{Explosion})}{+s}\,(+\vee\text{E.})^{1,2} \qquad [\ominus s]^4}{\bot}\,(\text{Rejection})}{\ominus\neg t}\,(\text{SR}_1)^3 \qquad \dfrac{+T\ulcorner\neg t\urcorner}{+\neg t}\,(\text{+T-OUT})}{\dfrac{\bot}{+s}\,(\text{SR}_2)^4}\,(\text{Rejection})$$

The method exemplified by this derivation consists in assuming the desired conclusion of a truth rule (here, $+\neg t$) in a restricted context and discharging it by applying the truth rule in the global proof context. The method generalizes: applications of (+T-OUT) can be 'moved outside' a restricted proof context in many cases. Specifically, if $+A$ is derivable from global premises (that is, not dependent on a dischargeable assumption), then any argument that would be valid, except that a truth rule is applied to $+A$, can be rewritten to a valid argument for the same conclusion by using the coordination principles as in the derivation above.[16]

The restrictions on rules like disjunction elimination are therefore much less restrictive than they appear to be at first sight. Whenever one uses a side premise such as $+T\ulcorner\neg t\urcorner$, one may apply truth rules freely to it. So proof by cases only fails in cases in which one must apply a truth rule to one of the disjuncts. There seem to be only two major cases where this happens. First, the proof of paradox from $+L \vee \neg L$, which *should* fail. Second, the derivation of the material biconditional $+A \leftrightarrow T\ulcorner A\urcorner$, which, as we argued above, *should* fail too.

The concept of truth is however employed in many areas of inquiry. It has been argued that deflationism cannot properly account for all these applications (Boyd 1983). Deflationists have responded to this at length (Williams 1988; Horwich 1998). Examination of these responses shows that they do not require full transparency—it suffices that A and $T\ulcorner A\urcorner$ can be inferred from one another. Hence, despite them not preserving evidence, our truth rules suffice to explain the various uses of truth.

The failure of full transparency allows us to treat Richard Kimberly Heck's (2012)

[16] Thus, the coordination principles play a role similar to the Cut rule in the sequent calculus. This points to a similarity between our approach to use restricted coordination principles* and David Ripley's (2013a; 2013b) proposal to avoid the semantic paradoxes by restricting Cut on the basis of a bilateralist interpretation of multiple-conclusion sequent calculi (in which a sequent $\Gamma \Rightarrow \Delta$ is read as expressing the incoherence of asserting all members of Γ and rejecting all members of Δ, see Restall 2005). The similarities do not go far, however: in Ripley's theory, the consequence relation is not transitive; it is in ours, since we *do* include the analogue of unrestricted Cut, the coordination principles. Ripley, for his part, retains unrestricted *reductio*. Rohan French (2016) has argued for failures of the structural rule of Reflexivity on the basis of a different bilateral interpretation of the sequent calculus: $\Gamma \Rightarrow \Delta$ expresses the incoherence of not rejecting all members of Γ and not asserting any member of Δ. Like Ripley, French forfeits a structural rule of the sequent calculus, but retains the meta-rule of *reductio*. Our theory validates (the analogue of) Reflexivity, but not *reductio*.

Differences run deeper still. Ours is a natural deduction calculus, we distinguish two signs for rejection and we treat asserted and rejected sentences as premises and conclusions of inferences, whereas sequent approaches use assertion and rejection to give a theory of logical consequence as the relation holding between antecedents and succedents of sequents. Bilateral sequent calculi have also been applied to validity paradoxes (Hlobil 2019; Rosenblatt 2021). We hope to explore applications of our approach to these paradoxes in future work.

strong Liar paradox exactly like the basic Liar paradox. Instead of using a *sentence* L that is equivalent to its own falsity, Heck considers the *term* λ that is a name for the formula expressing the falsity of λ. That is, $\lambda = \ulcorner \neg T(\lambda) \urcorner$. Although something stronger than the standard diagonal lemma is required to show that such terms exist, Heck argues that λ deserves our attention, since it is a more faithful formalization of the paraphrase *the sentence that says of itself that it is false*.

If one admits the existence of λ, one can derive a contradiction from "very meagre logical resources" (Heck 2012, 36). The following derivation adapts Heck's argument to the language of $\mathsf{WBL_T}$.[17]

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{+T(\lambda) \lor \neg T(\lambda)}{+T(\ulcorner \neg T(\lambda) \urcorner) \lor \neg T(\lambda)}\,\lambda = \ulcorner \neg T(\lambda) \urcorner}{+\neg T(\lambda) \lor \neg T(\lambda)}\,\text{Transparency}}{+\neg T(\lambda)}\,p \lor p \vdash p}{+T(\ulcorner \neg T(\lambda) \urcorner)}\,(\text{+T-IN})}{+T(\lambda)}\,\ulcorner \neg T(\lambda) \urcorner = \lambda \qquad \dfrac{\dfrac{\dfrac{+T(\lambda) \lor \neg T(\lambda)}{+T(\ulcorner \neg T(\lambda) \urcorner) \lor \neg T(\lambda)}\,\lambda = \ulcorner \neg T(\lambda) \urcorner}{+\neg T(\lambda) \lor \neg T(\lambda)}\,\text{Transparency}}{+\neg T(\lambda)}\,p \lor p \vdash p}{\bot}\,(\text{Law of Non-Contradiction})$$

The argument fails in $\mathsf{WBL_T}$, because $\mathsf{WBL_T}$ does not validate the required instances of Transparency. To eliminate truth predicates in a disjunction, that is to show that $+T(\ulcorner \neg T(\lambda) \urcorner) \lor \neg T(\lambda) \vdash +\neg T(\lambda) \lor \neg T(\lambda)$, one must apply (+T-OUT) within proof by cases. To wit:

$$\# \; \dfrac{+T(\ulcorner \neg T(\lambda) \urcorner) \lor \neg T(\lambda) \qquad \dfrac{\dfrac{[+T(\ulcorner \neg T(\lambda) \urcorner]^1}{+\neg T(\lambda)}\,(\text{+T-OUT})}{+\neg T(\lambda) \lor \neg T(\lambda)}\,(+\lor\text{I.}) \qquad \dfrac{[+\neg T(\lambda)]^2}{+\neg T(\lambda) \lor \neg T(\lambda)}\,(+\lor\text{I.})}{+\neg T(\lambda) \lor \neg T(\lambda)}\,(+\lor\text{E.})^{12}$$

But this derivation is invalid, since (+T-OUT) may not be applied in (+∨E.) and the method outlined above cannot be applied here. The failure of full transparency within disjunctive contexts saves $\mathsf{WBL_T}$ from the strong Liar paradox.

Heck anticipated the possibility of rejecting Transparency. They note that even if Transparency is not valid, it suffices for their arguments that the following is the case

[17] Heck presents a second argument using the same logical resources, so we omit it here.

for each sentence s.

(*) $$\neg(s \wedge T\ulcorner\neg s\urcorner)$$

Heck defends (*) by noting that even if truth is not transparent, asserting *snow is white and* ⌜*snow is not white*⌝ *is true* is unacceptable, hence one should accept (*). We agree with the observation, but not with the conclusion. $\mathsf{WBL_T}$ shows that $+s \wedge T\ulcorner\neg s\urcorner$ entails a contradiction. This explains the unacceptability of sentences such as *snow is white and* ⌜*snow is not white*⌝ *is true*. But since we use a truth rule to derive the contradiction, we cannot apply Smileian *reductio** to conclude $+\neg(s \wedge T\ulcorner\neg s\urcorner)$. Using $\mathsf{WBL_T}$, one can explain Heck's observation without having to concede (*).

6 Supervaluations, Quantifiers and Compositionality

Weak bilateral logic bears some similarities to supervaluationist logics. Notably, the meta-inferences that characteristically fail according to supervaluationism are exactly those that fail in WBL. We now demonstrate that these similarities run deep: the extension of the truth predicate defined by $\mathsf{WBL_T}$ can also be obtained by a supervaluational hierarchy. We use this result to outline a strategy for axiomatizing the extensions of the truth predicates defined by these hierarchies.

Kripke (1975) introduced the idea to construct the extension of the truth predicate with a stepwise procedure, beginning with no true sentences at all and continuing to add all further sentences whose truth is non-paradoxical in front of the sentences that have already been established to be in the extension of the truth predicate. The supervaluationial technique is one way of spelling out this broad plan. It is typically used to compute an extension of the truth predicate in True Arithmetic. This is the theory consisting of all sentences in the language $\mathcal{L}_A$ of arithmetic which are true in the standard model $\mathbb{N}$. True Arithmetic is a first-order theory, so we first extend $\mathsf{WBL_T}$ with quantifiers. The meaning-conferring rules for the universal quantifier are as follows.

$$(+\forall\text{I.})\ \frac{+A[a/x]}{+\forall x\,A}\ \text{if } a \text{ is any constant symbol not occurring in premises or undischarged assumptions used to derive } A[a/x] \qquad (+\forall\text{E.})\ \frac{+\forall x\,A}{+A[t/x]}$$

These are the standard rules for the universal quantifier in natural deduction, except that their premises and conclusions are positively signed. The existential quantifier can be defined as the dual of the universal one as usual. We use $\mathsf{QWBL_T}$ to denote the result of extending $\mathsf{WBL_T}$ with the universal quantifier rules.

In the context of True Arithmetic, one furthermore requires the *ω-Rule*, which allows one to infer a universally quantified sentence from all its instances for the natural numbers. As usual, boldface numerals are canonical names for the numbers.

$$(\omega\text{-Rule})\ \frac{+A(\mathbf{0}) \qquad +A(\mathbf{1}) \qquad +A(\mathbf{2}) \qquad \ldots}{+\forall n\,A}$$

Let $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$ be the result of extending $\mathsf{QWBL_T}$ with the ω-Rule. The logic of the truth predicate according to $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$ is the same as that of the truth predicate defined by the *van Fraassen hierarchy*. This hierarchy was suggested by Kripke (1975), adapting the supervaluation technique of van Fraassen (1971).

Let $\mathbb{N} \models A$ be defined in the usual way for sentences $A \in \mathcal{L}_A$. Given a set of numbers τ, one can extend the definition of $\models$ to the language of first-order arithmetic with a truth predicate $\mathcal{L}_{A_T}$ by letting $\mathbb{N}, \tau \models Tu$ just in case $u^{\mathbb{N}} \in \tau$ (where u is a term and $u^{\mathbb{N}}$ is its denotation in $\mathbb{N}$). By identifying names for sentences with canonical names for their Gödel numbers, it follows in particular that $\mathbb{N}, \tau \models T\ulcorner A \urcorner$ just in case $\ulcorner A \urcorner^{\mathbb{N}} \in \tau$. The van Fraassen hierarchy can then be recursively defined as follows.[18]

- Base: $\sigma_0 = \emptyset$.
- Successor: $\ulcorner A \urcorner^{\mathbb{N}} \in \sigma_{\alpha+1}$ iff for all sets $\tau \supseteq \sigma_\alpha$ such that

[18] This hierarchy is sometimes defined in different, non-equivalent ways. Field (2008) and Meadows (2015) supervaluate at the successor step over all $\tau \supseteq \sigma_\alpha$ with condition (i), whereas Oms (2020) supervaluates over all $\tau \supseteq \sigma_\alpha$ with condition (ii). These are not equivalent. Given some σ_α where $\ulcorner \neg A \urcorner^{\mathbb{N}} \in \sigma_\alpha$, Field and Meadows (but not Oms) permit that $\ulcorner A \urcorner^{\mathbb{N}} \in \tau$ while also $\ulcorner \neg A \urcorner^{\mathbb{N}} \in \tau \supseteq \sigma_\alpha$. This cannot happen under Kripke's (1975) original definition. Kripke supervaluates over τ that assign an extension and anti-extension to the truth predicate that are supersets of, respectively, the extension and anti-extension assigned in step α. Extension and anti-extension are disjoint. If $\ulcorner \neg A \urcorner^{\mathbb{N}}$ is in the extension of the truth predicate in step α, then $\ulcorner A \urcorner^{\mathbb{N}}$ is in its anti-extension at step α, hence also in the anti-extension assigned by every τ in the supervaluation, so $\ulcorner A \urcorner^{\mathbb{N}}$ cannot be in the extension assigned by any τ. Our definition is equivalent to Kripke's (as, less obviously, is Oms's). We assume that this is what is intended by everyone.

(i.) $\tau \cap \{\ulcorner \neg A \urcorner^{\mathbb{N}} \mid \ulcorner A \urcorner^{\mathbb{N}} \in \sigma_\alpha\} = \emptyset$ and

(ii.) $\tau \cap \{\ulcorner A \urcorner^{\mathbb{N}} \mid \ulcorner \neg A \urcorner^{\mathbb{N}} \in \sigma_\alpha\} = \emptyset$,

it is the case that $\mathbb{N}, \tau \models A$.

- Limit: if λ is a limit, let $\sigma_\lambda = \bigcup_{\alpha<\lambda} \sigma_\alpha$

It is easy to see that the hierarchy is non-decreasing, that is that for all $\beta < \alpha$, $\sigma_\beta \subseteq \sigma_\alpha$. Hence, by a fixed-point argument, there is some λ such that $\sigma_\lambda = \sigma_{\lambda+1}$. After this λ, nothing will change anymore, so we may regard σ_λ as *the* extension σ^{vF} of the truth predicate generated by the van Fraassen hierarchy. The sentences that are determined to be arithmetical truths by this hierarchy are exactly those sentences that $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$ proves from True Arithmetic. We can state this fact precisely by helping ourselves to some additional notation. Let $+\mathbf{TA}$ be the set of assertions of T-free arithmetical facts. In symbols: $+\mathbf{TA} = \{+X \mid \mathbb{N} \models X, X \in \mathcal{L}_A\}$. We have:

Theorem 6.1. *For all A in $\mathcal{L}_{A_T}$: $\ulcorner A \urcorner^{\mathbb{N}} \in \sigma^{vF}$ iff $+\mathbf{TA} \vdash^{\mathsf{QWBL}^{\omega}_{\mathsf{T}}} +A$.*

That is, $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$ corresponds to the internal logic of σ^{vF}. The proof is in the Appendix.

There are other proof theories for σ^{vF} (Meadows 2015; Stern 2018). Like $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$, which includes the ω-Rule, they incorporate infinitary components. This points to a different but related problem, namely that of stating a *recursive* set of axioms that characterizes a sufficiently large subset of σ^{vF}. The *external* logic in which these axioms are stated should be sound for the *internal* logic of truth defined by them (Halbach and Horsten 2005; 2006). That is, whenever A is a theorem in the external logic, $T\ulcorner A \urcorner$ should follow from the axioms, for otherwise the external logic is unsound in that it proves something untrue.

For example, Solomon Feferman's (1991) theory KF axiomatizes a fragment of the Strong Kleene version of Kripke's theory of truth, but KF itself is phrased in classical logic. There are theorems of classical logic whose truth does not follow from KF. Halbach and Horsten (2006) succeed in developing the theory PKF that axiomatizes a fragment of the Strong Kleene version of Kripke's theory of truth and is sound for its external logic. Something like PKF has not yet been found for the supervaluational approaches. Horsten (2011, 139) states that "it is unclear how a natural formalization of supervaluation fixed points would look".

We are now in a position to address this problem for the van Fraassen hierarchy. Using QWBL as the external logic, we can provide a natural, recursive axiomatization of the van Fraassen supervaluational theory of truth. More precisely, we add to $\mathsf{QWBL_T}$ all formulae $+P$ where P is an axiom of Peano Arithmetic (including the induction scheme over the language including T). It is trivially the case that the external logic is sound for the internal logic, that is when $+A$ is a theorem, then $+T\ulcorner A\urcorner$ is a theorem too. That is, all theorems are truths. It is a corollary of Theorem 6.1 that the truths form a subset of σ^{vF}. In the Appendix, we provide a model theory for which $\mathsf{QWBL_T}$ is sound and complete.

We can consider further principles to axiomatize other supervaluational hierarchies. A notable case is the hierarchy defined by Cantini (1990), in which the $\tau \supseteq \sigma$ in the successor step must be such that for all sentences A, if $\ulcorner A\urcorner^{\mathbb{N}} \in \tau$, then $\ulcorner \neg A\urcorner^{\mathbb{N}} \notin \tau$. Let σ^C be the fixed point of this hierarchy and $\mathsf{QWBL^{\omega}_{TC}}$ be the result of extending $\mathsf{QWBL^{\omega}_{T}}$ with the following inference rule (C), where it is permitted to use (C) in Smileian *reductio**.

$$(\mathsf{C})\ \frac{+T\ulcorner\neg A\urcorner}{+\neg T\ulcorner A\urcorner}$$

The rule (C) is already derivable in $\mathsf{QWBL^{\omega}_{T}}$ by using truth rules, but allowing its use in Smileian *reductio** strengthens the calculus so that the truths derivable from true arithmetic in $\mathsf{QWBL^{\omega}_{TC}}$ are exactly the members of σ^C.

Theorem 6.2. *For all $A \in \mathcal{L}_{A_T}$: $\ulcorner A\urcorner^{\mathbb{N}} \in \sigma^{\mathsf{C}}$ iff $+\mathbf{TA} \vdash^{\mathsf{QWBL^{\omega}_{TC}}} +A$.*

Cantini (1990) provides an axiomatization VF of a fragment of σ^C, but phrases it in classical logic, which has theorems that are not members of σ^C. Halbach (2011, 266) writes that he is "not aware of any attempt to find a system in supervaluational logic that relates to VF in the way to PKF relates to KF". Theorem 6.2 can be used to provide such a system. We can obtain an axiomatization of a fragment of σ^C with QWBL as the external logic by adding the Peano Axioms to $\mathsf{QWBL_{TC}}$.

Field (2008, ch. 11) criticizes these supervaluational hierarchies for failing to define a *compositional* extension of the truth predicate. Halbach and Horsten (2005, 207) similarly demand that "the truth predicate should commute with quantifiers and connectives". For example, although it is the case that $\ulcorner A \wedge B \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{vF}}$ if and only if $\ulcorner A \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{vF}}$ and $\ulcorner B \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{vF}}$, it is *not* the case that the biconditional $\ulcorner T \ulcorner A \wedge B \urcorner \leftrightarrow (T \ulcorner A \urcorner \wedge T \ulcorner B \urcorner) \urcorner^{\mathbb{N}}$ is a member of σ^{vF} (and the same for σ^{C}). The most dramatic failure of compositionality occurs with the quantifiers. One may attempt to phrase compositionality as follows.

$$\text{(QC)} \quad +\forall \mathbf{n}.T \ulcorner A(\mathbf{n}) \urcorner \leftrightarrow T \ulcorner \forall n\, A \urcorner.$$

There are formal problems with quantifying into quotational position, so the left-hand-side quantifier in (*) substitutionally quantifies over names (Halbach and Horsten 2005).[19]

In theories where not every member of the domain has a name, the range of the substitutional quantifier is limited, so one may not expect (QC) to be true. But in the context of True Arithmetic, where every number is denoted by its canonical name, it appears reasonable to demand that (QC) be valid. However, the left-to-right direction of (QC) cannot be true in any standard model of arithmetic.[20]

When approaching the hierarchies from the perspective of $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$, it becomes clear that requiring compositionality to be expressed by *material* biconditionals like (QC) is too high a demand. The inference rules of $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$ define the compositional meaning of the truth predicate, but this definition is not expressible in terms of material conditionals. This is because inference in $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$ need not preserve evidence, whereas given a material conditional $A \to B$, to infer B from A does preserve evidence. Thus, while it is the case that according to $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$, the truth of a conjunction is equivalent to the truth of both conjuncts, this is *not* a material equivalence. So it is to be expected that the corresponding material biconditional is not a truth.

The worry about compositionality may not have been entirely put to rest. One may

[19] Formally, one takes $\forall \mathbf{n}.T \ulcorner A(\mathbf{n}) \urcorner$ to abbreviate $\forall x\, \varphi^{A}(x) \to Tx$, where $\mathbb{N} \models \varphi^{A}(x)$ just in case x is the Gödel number of any sentence obtained by replacing the variable n in A by a closed term.

[20] We omit a formal a proof, since Field (2008, 185) gives all the necessary details, crediting Vann McGee (1992) with the crucial observation.

want to to state compositionality using *object language sentences*. Roughly, although truth in $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$ is compositional, it does not follow that *truth is compositional* is true because there are no object language sentences expressing compositionality. To phrase such sentences, we extend $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$ with a non-material conditional $\Rightarrow$ that is the embeddable version of inference in $\mathsf{QWBL}_{\mathsf{T}}$ and that can be read as *if A, then it follows that B*. This conditional is defined by the inference rules of *modus ponens* and *unrestricted* conditional proof.

$$(+\Rightarrow\text{I.})\ \frac{\begin{array}{c}[+A]\\ \vdots\\ +B\end{array}}{+A \Rightarrow B} \qquad (+\Rightarrow\text{E.})\ \frac{+A \qquad +A \Rightarrow B}{+B}$$

This is the commitment-preserving conditional: in uttering $+A \Rightarrow B$, one asserts that the inference from A to B preserves commitment. This inference need not preserve evidence, however. This means that, since $(+\Rightarrow\text{E.})$ allows one to infer $+B$ from $+A$ on the grounds that $+A \Rightarrow B$, applications of this rule do not preserve evidence. We must therefore exclude $(+\Rightarrow\text{E.})$ from Smileian *reductio**.

Having extended $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$ with these rules and restricted Smileian *reductio** appropriately, one can schematically derive object language sentences expressing compositionality. For arbitrary sentences A:

$$+T\ulcorner\neg A\urcorner \Leftrightarrow \neg T\ulcorner A\urcorner,$$

$$+T\ulcorner A \wedge B\urcorner \Leftrightarrow T\ulcorner A\urcorner \wedge T\ulcorner B\urcorner,$$

$$+\forall \mathbf{n}.T\ulcorner A(\mathbf{n})\urcorner \Leftrightarrow T\ulcorner \forall n\, A\urcorner.$$

This appears to be sufficient to address Field's worries, but Halbach and Horsten (2005) have a stronger notion of compositionality in mind. They demand universally quantified compositional principles like the following for negation. (The predicate `Sent` denotes the property of being the Gödel number of a sentence and we adopt the convention that square brackets indicate operations in the coding language, for example $[\neg n]$ is the Gödel number of the negation of the sentence whose Gödel number is n.)

$$\forall n(\texttt{Sent}(n) \to (T[\neg n] \leftrightarrow \neg Tn)).$$

Using the ω-Rule we can immediately derive such universal generalizations from the schematic results above. For instance, for negation we can derive the following (and analogously for the other logical constants).

$$+\forall n(\texttt{Sent}(n) \to (T[\neg n] \Leftrightarrow \neg Tn)).$$

Absent the ω-Rule, however, this universal generalization does not follow from the schematic derivability of $+T\ulcorner\neg A\urcorner \Leftrightarrow \neg T\ulcorner A\urcorner$ (and likewise for $\wedge$ and $\forall$). Halbach and Horsten do not appear to consider infinitary principles like the ω-Rule to be acceptable means to satisfy their requirement.

Again using the commitment-preserving conditional, we can state a recursive definition of truth in QWBL that is compositional in the sense of Halbach and Horsten as follows. ($\texttt{At}$ denotes the property of being the Gödel number of an atomic sentence in the language of arithmetic, and $\texttt{T}_0$ is the definable truth predicate for atomic sentences).

1. $+\forall n\, \texttt{At}(n) \to (Tn \Leftrightarrow T_0 n)$.
2. $+\forall n\, \texttt{Sent}(n) \to (T[\neg n] \Leftrightarrow \neg Tn)$
3. $+\forall n \forall m\, \texttt{Sent}(n) \wedge \texttt{Sent}(m) \to (T[n \wedge m] \Leftrightarrow (Tn \wedge Tm))$
4. $+\forall n\, \texttt{Sent}(n) \to (\forall \mathbf{m}\, Tn(\mathbf{m}) \Leftrightarrow T[\forall m\, n])$.

These axioms are all derivable from True Arithmetic in $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$, so this axiomatic theory of truth also broadly corresponds to a fragment of a fixed point. A more precise assessment is hindered by the fact that the supervaluational technique cannot be applied to sentences with commitment-preserving conditionals. This is because we have not provided a definition of when $\mathbb{N} \models p \Leftrightarrow q$.[21] We leave a deeper investigation of the axiomatic theory phrased using the commitment-preserving conditional to further work.

[21] We doubt that such a definition can be found, as a natural model theory for $\Leftrightarrow$ treats it as a modal operator scoping over the individual models forming the *QWBL$_T$ models* defined in the Appendix.

It is nonetheless possible to extend $\mathsf{QWBL}^\omega_\mathsf{T}$ so that the material biconditionals stating the compositionality of conjunction and negation become derivable (although the quantifier case remains hopeless). Doing so will establish a connection between a motivated extension of $\mathsf{QWBL}^\omega_\mathsf{T}$ and another supervaluational hierarchy. The idea is to add rules stating that the logic under $+T$ and $-T$ is the same as the logic of $+$ and $-$. Formally, let $\mathsf{QWBL}^\omega_\mathsf{TM}$ be the extension of $\mathsf{QWBL}^\omega_\mathsf{T}$ with the following rules for material compositionality. For readability, we write these as sequent rules.[22]

$$(+\mathrm{MC})\ \frac{+A_1, +A_2, ..., +A_n \ \vdash\ \bot}{+T\ulcorner A_1\urcorner, +T\ulcorner A_2\urcorner, ..., +T\ulcorner A_n\urcorner \ \vdash\ \bot}\ \begin{array}{l}\text{if the inference to } \bot \text{ uses no premises}\\ \text{signed with } \ominus \text{ and no truth rules.}\end{array}$$

$$(-\mathrm{MC})\ \frac{-A_1, -A_2, ..., -A_n \ \vdash\ \bot}{-T\ulcorner A_1\urcorner, -T\ulcorner A_2\urcorner, ..., -T\ulcorner A_n\urcorner \ \vdash\ \bot}\ \begin{array}{l}\text{if the inference to } \bot \text{ uses no premises}\\ \text{signed with } \ominus \text{ and no truth rules.}\end{array}$$

Both (+MC) and (–MC) are already derivable in $\mathsf{QWBL}^\omega_\mathsf{T}$, but their derivation uses the truth rules, which prevents their application in Smileian *reductio**. By using (+MC) and (–MC) in Smileian *reductio**, one can prove the material compositionality of the propositional connectives.

Theorem 6.3. *The following are derivable in* $\mathsf{QWBL}^\omega_\mathsf{TM}$.

- $+T\ulcorner\neg A\urcorner \leftrightarrow \neg T\ulcorner A\urcorner$
- $+T\ulcorner A \wedge B\urcorner \leftrightarrow (T\ulcorner A\urcorner \wedge T\ulcorner B\urcorner)$.

The proof is in the Appendix. This theorem is compatible with our arguments that the truth rules do not preserve evidence. Our counterexample to the evidence preservation of (+T-OUT) is a situation in which we have evidence for Thomas having written a true sentence W on his whiteboard: we may have evidence for $\ulcorner W\urcorner$ *is true*, but no evidence for W itself. Now, in addition to knowing that Thomas wrote down a true sentence, we may know some structural properties of that sentence. If we have evidence that he wrote a true conjunction we may let $\ulcorner Q\urcorner$ and $\ulcorner R\urcorner$ be names for its conjuncts. Then, evidence for $\ulcorner W\urcorner$ *is true* is indeed also evidence for $\ulcorner Q\urcorner$ *is true and* $\ulcorner R\urcorner$ *is true* (but not

[22] We may also code these inferences directly into Smileian *reductio** by weakening its restriction to disallow the use truth rules *except* when one applies (+T-OUT) to every premise or (–T-OUT) to every premise. We think this is the most principled, albeit cumbersome, way of establishing material compositionality.

necessarily for W, Q or R). Similarly, if we know that Thomas wrote a true negation and let $\ulcorner Q \urcorner$ be a name for the sentence whose negation is on Thomas's whiteboard, we have evidence for $\ulcorner Q \urcorner$ *is not true*.[23]

$\mathsf{QWBL}^{\omega}_{\mathsf{TM}}$ corresponds to the internal logic of the truth predicate defined by the supervaluational hierarchy first considered by Kripke (1975) and obtained by using $\tau \supseteq \sigma$ in the successor step that are maximally classically consistent sets of sentences, that is that have the following properties.[24]

(i.) τ is classically consistent (that is the set of sentences whose Gödel numbers are in τ is classically consistent); and

(ii.) for any A, either $\ulcorner A \urcorner^{\mathbb{N}} \in \tau$ or $\ulcorner \neg A \urcorner^{\mathbb{N}} \in \tau$.

Again, it is easy to see that the hierarchy is non-decreasing. So there is a fixed point σ^{mc}. The members of σ^{mc} are exactly the truths that $\mathsf{QWBL}^{\omega}_{\mathsf{TM}}$ derives from True Arithmetic.

Theorem 6.4. *For all $A \in \mathcal{L}_{A_T}$: $A \in \sigma^{mc}$ iff $+\mathbf{TA} \vdash^{\mathsf{QWBL}^{\omega}_{\mathsf{TM}}} +A$.*

So, $\mathsf{QWBL}^{\omega}_{\mathsf{TM}}$ is the logic of σ^{mc}. One can now recursively axiomatize fragments of σ^{mc} as we have done above for the other fixed points: by taking $\mathsf{QWBL}_{\mathsf{TM}}$ over the asserted Peano Axioms.

7 Conclusion: Deflationism and Supervaluations

Tarski's Theorem is usually taken to show that any language containing enough arithmetic cannot express its own truth predicate. This interpretation presupposes that 'ex-

[23] This may explain why the quantifiers resist material compositionality. Halbach and Horsten (2005, 208) note that "strictly speaking", (QC) does not express the compositionality of the quantifier, because the substitution instances $A(\mathbf{n})$ are not literally parts of $\forall x\, A$ (that is they are not subformulae). Now, the reason why the inference from $\ulcorner Q \wedge R \urcorner$ *is true* to $\ulcorner Q \urcorner$ *is true and* $\ulcorner R \urcorner$ *is true* can be said to preserve evidence is that we can exploit knowledge about how the sentence on Thomas's whiteboard decomposes into parts. But, taking up Halbach and Horsten's observation, we do not literally have knowledge about the parts of a universally quantified sentence when the sentence on Thomas's whiteboard is a substitutional scheme. So, perhaps, $\forall \mathbf{n}.T \ulcorner A(n) \urcorner \to T \ulcorner \forall n\, A \urcorner$ cannot be validated as we do not have structural knowledge about the parts of the formula in the consequent. To put flesh on these bones, however, one would need to look more closely into the relationship between the formal device $\forall \mathbf{n}$, quotation, and evidence. We leave this to future work.

[24] Kripke (1975, 711) defines consistency not as in (i), but as τ not containing both $\ulcorner A \urcorner^{\mathbb{N}}$ and $\ulcorner \neg A \urcorner^{\mathbb{N}}$ for any A. This leaves open that some τ are *classically* inconsistent because, say, they contain both $\ulcorner A \wedge A \urcorner^{\mathbb{N}}$ and $\ulcorner \neg A \urcorner^{\mathbb{N}}$. Our clause (i) excludes such deviant cases; this is how Kripke is typically interpreted (see Field 2008, 180).

pressing one's own truth predicate' means validating all *material* biconditionals of the disquotational scheme, that is $A \leftrightarrow T\ulcorner A \urcorner$ for all sentences A. We have seen that this presupposition is too restrictive. Our new diagnosis of the problem is that the truth rules do not preserve evidence and so should not be stated as material conditionals, which we backed up with an independent argument.

Inferential deflationism breaks free from the confines of Tarski's result by formulating disquotation via inferences that preserve commitment, but not evidence, and hence cannot be stated as material conditionals. The same diagnosis and strategy can be applied to the strong Liar. Formulating disquotation in terms of rules relating truth to assertion (and rejection) moreover allows the inferential deflationist to make good on the claim that the truth predicate is a vehicle for endorsement (and opposition) without running into the Frege-Geach problem.

Horwich (1998) initially sought to free truth from the limitations imposed by Tarski by accepting only the largest possible set of instances of the disquotational scheme that does not entail paradox. McGee (1992) demonstrated that this leaves the theory of truth vastly underspecified. In response, Horwich (2010) proposed an iterative construction that determines a maximal set of *ground* truths for which one can accept material disquotation. A recent result by Sergi Oms (2020), adapting a formalization of Horwich's strategy by Thomas Schindler (2020), shows that the set of truths obtained by iteratively taking ground instances is exactly σ^{vF}, the set of truths defined by the van Fraassen hierarchy. Our result that $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$ axiomatizes the internal logic of σ^{vF} independently confirms a close relationship between this hierarchy and deflationism.

This helps assuage a worry raised by Meadows (2013, 230). He challenges the supervaluational approach on the grounds that there is no "principled reason to choose between" the different hierarchies. Absent any reason to think of one hierarchy as more natural than the others, *no* hierarchy can be said to produce *the* set of truths over arithmetic. The above results suggest that the van Fraassen hierarchy is a particularly natural choice from a deflationist perspective, being obtained via two independent methods of spelling out the deflationist project.

However, considerations other than naturalness may be relevant for the choice of the right hierarchy. For example, one may desire to fulfill Horwich's original ambition of obtaining a maximal set of truths. A better candidate for maximality than the van Fraassen hierarchy is the maximally consistent hierarchy: its fixed point σ^{mc} contains many more truths than σ^{vF}. Although Horwich (2010, 92) claims that the supervaluational hierarchies do not "square with minimalism", Oms's result shows how one can get to σ^{vF} with tools acceptable to the minimalist. If our inferential deflationism and the sequent rules for material compositionality are acceptable as well (as we think they are), minimalism can go up to σ^{mc}. We leave it as an open question whether it is possible to go beyond the maximally consistent hierarchy, as characterized by the inferential deflationist calculus $\mathsf{QWBL}^{\omega}_{\mathsf{TM}}$, or whether we have achieved maximality in minimalism.

Appendix: Proofs

Truth and Evidence

A *WBL model* is a countably infinite set $\mathcal{M}$ of classical valuations (Incurvati and Schlöder 2017). A *WBL_T model* is a tuple $(\mathcal{M}, t)$ where $\mathcal{M}$ is a WBL model and $t : \mathcal{M} \to \mathcal{M}$ is a bijection. Then define:

- For $M \in \mathcal{M}$ and a bijection t, define *local satisfaction* as $M, t \Vdash p$ iff$_{\mathrm{Def}}$ $M(p) = 1$ for propositional atoms p; $M, t \Vdash \neg A$ iff$_{\mathrm{Def}}$ $M, t \nVdash A$; $M, t \Vdash A \wedge B$ iff$_{\mathrm{Def}}$ $M, t \Vdash A$ and $M, t \Vdash B$; and $M, t \Vdash T\ulcorner A \urcorner$ iff$_{\mathrm{Def}}$ $t(M), t \Vdash A$.
- Then define *global satisfaction* as $\mathcal{M}, t \models +A$ iff$_{\mathrm{Def}}$ for all $M \in \mathcal{M}$, $M, t \Vdash A$; $\mathcal{M}, t \models -A$ iff$_{\mathrm{Def}}$ for all $M \in \mathcal{M}$, $M, t \nVdash A$; and $\mathcal{M}, t \models \ominus A$ iff$_{\mathrm{Def}}$ there is some $M \in \mathcal{M}$ with $M, t \nVdash A$.

$\mathsf{WBL_T}$ is sound for the class of $\mathsf{WBL_T}$ models. It is straightforward to prove that our truth rules are sound and the soundness argument for WBL of Incurvati and Schlöder

(2017) shows the rest. The Liar biconditional $+L \leftrightarrow \neg T \ulcorner L \urcorner$ is satisfied for a propositional atom L in a model $(\mathcal{M}, t)$ where for all $M \in \mathcal{M}$: $M(L) = 1$ iff $t(M)(L) = 0$. One such model is $\mathcal{M} = \{M_i \mid i \in \omega\}$ where $M_i(L) = 1$ iff i is even, $t(M_i) = M_{i+1}$ when i is even and $t(M_i) = M_{i-1}$ when i is odd.

Supervaluations, Quantifiers and Compositionality

The van Fraassen hierarchy. We prove the following theorem.

Theorem 6.1. For all A in $\mathcal{L}_{A_T}$: $\ulcorner A \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{vF}}$ iff $+\mathbf{TA} \vdash^{\mathsf{QWBL}^{\omega}_{\mathsf{T}}} +A$.

Throughout the proof we write $\vdash$ for $\vdash^{\mathsf{QWBL}^{\omega}_{\mathsf{T}}}$. Recall that σ^{vF} is the fixed point of the hierarchy supervaluating over τ with:

(i.) $\tau \cap \{\ulcorner \neg A \urcorner^{\mathbb{N}} \mid \ulcorner A \urcorner^{\mathbb{N}} \in \sigma_\alpha\} = \emptyset$ and

(ii.) $\tau \cap \{\ulcorner A \urcorner^{\mathbb{N}} \mid \ulcorner \neg A \urcorner^{\mathbb{N}} \in \sigma_\alpha\} = \emptyset$,

The right-to-left direction of Theorem 6.1 follows from the result that the consequence relation of $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$ preserves membership in σ^{vF} in the following sense. Define $+A \,\tilde{\in}\, \sigma^{\mathrm{vF}}$ iff $\ulcorner A \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{vF}}$, $-A \,\tilde{\in}\, \sigma^{\mathrm{vF}}$ iff $\ulcorner \neg A \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{vF}}$ and $\ominus A \,\tilde{\in}\, \sigma^{\mathrm{vF}}$ iff $\ulcorner A \urcorner^{\mathbb{N}} \notin \sigma^{\mathrm{vF}}$. Then:

Theorem 7.1. *Let Γ be a set of signed formulae and φ be a signed formula. If for all $\psi \in \Gamma$, $\varphi \,\tilde{\in}\, \sigma^{vF}$ and $\Gamma \vdash \varphi$, then also $\varphi \,\tilde{\in}\, \sigma^{vF}$.*

Proof. It is straightforward to see that the rules for conjunction, negation and universal quantification in $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$ preserve membership in σ^{vF}. All instances of Smileian *reductio** are instances of classically valid inferences (see Incurvati and Schlöder 2017), so they preserve membership as well. Since the membership conditions for $+A$ and $-A$ are contrary, (Strong Rejection) also preserves membership. Similarly, since the membership conditions for $+A$ and $\ominus A$ are contradictory, the coordination principles preserve membership. Hence it suffices to verify that the truth rules preserve membership in σ^{vF}.

- (+T-OUT). Suppose that $+T \ulcorner A \urcorner \,\tilde{\in}\, \sigma^{\mathrm{vF}}$, so $\ulcorner T \ulcorner A \urcorner \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{vF}}$. Because σ^{vF} is a fixed point, this means that for all $\tau \supseteq \sigma^{\mathrm{vF}}$ with (i) and (ii) as in the successor step of the van Fraassen hierarchy, $\mathbb{N}, \tau \models T\ulcorner A \urcorner$. This means that for all such τ, $\ulcorner A \urcorner^{\mathbb{N}} \in \tau$. If it were the case that $\ulcorner A \urcorner^{\mathbb{N}} \notin \sigma^{\mathrm{vF}}$, one could find a τ with (i) and (ii) of which $\ulcorner A \urcorner^{\mathbb{N}}$ is not a member. As there is no such τ, $\ulcorner A \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{vF}}$. Thus $+A \mathrel{\tilde{\in}} \sigma^{\mathrm{vF}}$.

- (+T-IN). Suppose that $+A \mathrel{\tilde{\in}} \sigma^{\mathrm{vF}}$, that is $\ulcorner A \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{vF}}$. So all supersets τ of σ^{vF} contain $\ulcorner A \urcorner^{\mathbb{N}}$. So for all such τ, we have that $\mathbb{N}, \tau \models T\ulcorner A \urcorner$. This goes in particular for supersets τ with (i) and (ii), so by the fixed point property, $\ulcorner T\ulcorner A \urcorner \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{vF}}$, so $+T\ulcorner A \urcorner \mathrel{\tilde{\in}} \sigma^{\mathrm{vF}}$.
- (–T-OUT). Suppose that $-T\ulcorner A \urcorner \mathrel{\tilde{\in}} \sigma^{\mathrm{vF}}$, so $\ulcorner \neg T\ulcorner A \urcorner \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{vF}}$. Because σ^{vF} is a fixed point, this means that for all $\tau \supseteq \sigma^{\mathrm{vF}}$ with (i) and (ii), $\mathbb{N}, \tau \models \neg T\ulcorner A \urcorner$. This means that for all such τ, $\ulcorner A \urcorner^{\mathbb{N}} \notin \tau$. If it were the case that $\ulcorner \neg A \urcorner^{\mathbb{N}} \notin \sigma^{\mathrm{vF}}$, one could find a τ with (i) and (ii) of which $\ulcorner A \urcorner^{\mathbb{N}}$ is a member. As there is no such τ, $\ulcorner \neg A \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{vF}}$. Thus $-A \mathrel{\tilde{\in}} \sigma^{\mathrm{vF}}$.
- (–T-IN). Suppose that $-A \mathrel{\tilde{\in}} \sigma^{\mathrm{vF}}$, that is $\ulcorner \neg A \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{vF}}$. Consider any $\tau \supseteq \sigma^{\mathrm{vF}}$ with (i) and (ii). By assumption, $\ulcorner \neg A \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{vF}} \subseteq \tau$ and so by condition (ii), $\ulcorner A \urcorner^{\mathbb{N}} \notin \tau$. Thus $\mathbb{N}, \tau \not\models T\ulcorner A \urcorner$, that is $\mathbb{N}, \tau \models \neg T\ulcorner A \urcorner$. This goes for all the τ, so by the fixed point property, $\ulcorner \neg T\ulcorner A \urcorner \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{vF}}$, so $-T\ulcorner A \urcorner \mathrel{\tilde{\in}} \sigma^{\mathrm{vF}}$. □

Now, because for all $A \in \mathcal{L}_A$ with $\mathbb{N} \models A$ it is the case that $+A \mathrel{\tilde{\in}} \sigma^{\mathrm{vF}}$, this result entails that if $+\mathbf{TA} \vdash +B$, then $\ulcorner B \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{vF}}$. This is the right-to-left direction of Theorem 6.1.

The left-to-right direction of Theorem 6.1 follows from a model existence result. We say that a set of formulae Γ is *consistent** if it is not possible to derive $\bot$ from Γ by only evidence-preserving inferences. And we say that Γ is *maximally consistent** if it is consistent* and for all A, either $+A \in \Gamma$ or $+\neg A \in \Gamma$.

Lemma 7.2. *Every consistent* set Γ of $\mathcal{L}_{A_T}$-formulae has a maximally consistent* superset $\hat{\Gamma}$.*

Since $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$ is not compact, we must adjust the standard construction to take the ω-Rule into account.

Proof. Let Γ be a consistent* set of $\mathcal{L}_{A_T}$-formulae and write $\vdash^*$ for the evidence-preserving fragment of the consequence relation of $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$. Let $\{A_n \mid n \in \mathbb{N}\}$ be an enumeration of the sentences of $\mathcal{L}_{A_T}$. We define a sequence Γ_n, $n \in \mathbb{N}$ by recursion.

$$\Gamma_0 = \Gamma.$$

$$\Gamma_{n+1} = \begin{cases} \Gamma_n \cup \{+A_n\} & \text{if (a) } \Gamma_n \vdash^* +A_n, \\ \Gamma_n \cup \{+\neg A_n\} & \text{if (b) } \Gamma_n \nvdash^* +A_n \text{ and } A_n \text{ is not of the form } \forall x B, \\ \Gamma_n \cup \{+\neg A_n, +\neg B[\mathbf{k}/x]\} & \text{if (c) } \Gamma_n \nvdash^* +A_n \text{ and } A_n \text{ is of the form } \forall x B. \\ \quad \text{with } k \text{ minimal such} & \\ \quad \text{that } \Gamma_n \nvdash^* +B[\mathbf{k}/x] & \end{cases}$$

Note that a k as required in (c) always exists. If there were no such k, $\Gamma_n \vdash^* +B[\mathbf{k}/x]$ for all $k \in \mathbb{N}$ and so by the ω-Rule, $\Gamma_n \vdash^* +A_n$. But then we are in case (a).

Now let $\hat{\Gamma} = \bigcup_{n \in \mathbb{N}} \Gamma_n$. Clearly, $\hat{\Gamma}$ extends Γ. Moreover, $\hat{\Gamma}$ is maximal, since for each n it contains either $+A_n$ or $+\neg A_n$. It remains to show that $\hat{\Gamma}$ is consistent*.

We first show that all Γ_n are consistent*. By assumption, Γ_0 is consistent*. Towards a contradiction, let m be the least n such that $\Gamma_{n+1} \vdash^* \bot$. If in the construction of Γ_{m+1} we are in case (a), then $\Gamma_m \cup \{+A_m\} \vdash^* \bot$ and $\Gamma_m \vdash^* +A_m$, so Γ_m is already inconsistent*, contradicting the choice of m as minimal. If we are in case (b), then $\Gamma_m \cup \{+\neg A_m\} \vdash^* \bot$. By Bilateral *Reductio*, $\Gamma_m \vdash^* +A_m$, contradicting the fact that we are in case (b). Finally, if we are in case (c), then $\Gamma_m \cup \{+\neg\forall x B, +\neg B[\mathbf{k}/x]\} \vdash^* \bot$, so by Bilateral *Reductio* and the De Morgan laws, $\Gamma_m \vdash^* +\forall x B \vee B[\mathbf{k}/x]$. By (+∀E.) and disjunctive syllogism, it follows that $\Gamma_m \vdash^* +B[\mathbf{k}/x] \vee B[\mathbf{k}/x]$, so $\Gamma_m \vdash^* +B[\mathbf{k}/x]$, contradicting the choice of k. We reach a contradiction in all cases.

Thus all Γ_n are consistent*. To show that $\hat{\Gamma}$ is consistent*, it suffices to prove that for all formulae φ, if $\hat{\Gamma} \vdash^* \varphi$, then there is an n such that $\Gamma_n \vdash^* \varphi$. The proof is by induction. We only cover the inductive step for the ω-Rule, since all other arguments are standard.

So suppose that $\hat{\Gamma} \vdash^* +\forall x B$ by a derivation whose last step is the ω-Rule. This means that for every $k \in \mathbb{N}$, there is a shorter derivation showing that $\hat{\Gamma} \vdash^* +B[\mathbf{k}/x]$. By the induction hypothesis, it follows that for each $k \in \mathbb{N}$ there is an n_k such that $\Gamma_{n_k} \vdash^* +B[\mathbf{k}/x]$. Let m be such that $+\forall x B = +A_m$. Consider the construction of Γ_{m+1}.

If we are in case (a), then $\Gamma_m \vdash^* +\forall x B$, and we are done. We cannot be in case (b) by the form of A_m. If we are in case (c), $\Gamma_{m+1} = \Gamma_m \cup \{+\neg\forall x B, +\neg B[\mathbf{k}/x]\}$ for k minimal with $\Gamma_m \nvdash^* +B[\mathbf{k}/x]$. Consider n_k. If $n_k \leq m$, then $\Gamma_{n_k} \subseteq \Gamma_m$, so $\Gamma_m \vdash +B[\mathbf{k}/x]$, contradicting the choice of k. If $n_k > m$, then $\Gamma_{n_k} \supseteq \Gamma_m$, so $\Gamma_{n_k} \vdash^* +B[\mathbf{k}/x]$ and $\Gamma_{n_k} \vdash^* +\neg B[\mathbf{k}/x]$, contradicting the fact that Γ_{n_k} is consistent*. So we cannot be in case (c). □

We next show that every maximally consistent* extension of True Arithmetic has a model.

Theorem 7.3. *Let $\Gamma \supseteq +\mathbf{TA}$ be maximally consistent* and define $\tau = \{u^{\mathbb{N}} \mid +Tu \in \Gamma\}$. Then for all $+A \in \Gamma$, it is the case that $\mathbb{N}, \tau \models A$.*

Proof. We prove by induction on the construction of A that $+A \in \Gamma$ iff $\mathbb{N}, \tau \models A$.

- Suppose A is a T-free atom. True arithmetic decides all such A, so $+A \in \Gamma$ iff $\mathbb{N} \models A$ iff $\mathbb{N}, \tau \models A$.

- Suppose $A = Tu$. $+Tu \in \Gamma$ iff $u^{\mathbb{N}} \in \tau$ iff $\mathbb{N}, \tau \models Tu$.

- Suppose $A = B \wedge C$. First suppose that $+A \in \Gamma$. Since Γ is maximally consistent*, $+B \in \Gamma$ and $+C \in \Gamma$. By the induction hypothesis, $\mathbb{N}, \tau \models B$ and $\mathbb{N}, \tau \models C$, so $\mathbb{N}, \tau \models A$. The backward direction is analogous.

- Suppose $A = \neg B$. If $+\neg B \in \Gamma$ then because Γ is consistent*, $+B \notin \Gamma$, so by the induction hypothesis, $\mathbb{N}, \tau \not\models B$. So $\mathbb{N}, \tau \models \neg B$. Conversely, if $\mathbb{N}, \tau \models \neg B$, then $\mathbb{N}, \tau \not\models B$, so by the induction hypothesis $+B \notin \Gamma$. As Γ is maximally consistent*, $+\neg B \in \Gamma$.

- Suppose $A = \forall x\, B$. If $+\forall x\, B \in \Gamma$ then because Γ is maximally consistent*, for every natural number $\mathbf{n}$, $B[\mathbf{n}/x] \in \Gamma$. By the induction hypothesis, $\mathbb{N}, \tau \models B[\mathbf{n}/x]$. As this goes for any $\mathbf{n}$ and $\mathbb{N}$ is the standard model, $\mathbb{N}, \tau \models \forall x\, B$. Conversely, if $\mathbb{N}, \tau \models \forall x\, B$, then for all numbers $\mathbf{n}$, $\mathbb{N}, \tau \models B[\mathbf{n}/x]$, so for all $\mathbf{n}$, $B[\mathbf{n}/x] \in \Gamma$ by the induction hypothesis. Because Γ is maximally consistent* and $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$ contains the ω-Rule, $\forall x\, B \in \Gamma$. □

We are now in a position to prove the left-to-right direction of Theorem 6.1.

Proof. We first show that if $\ulcorner A \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{vF}}$, then $+\mathbf{TA} \vdash +A$ by an induction on the supervaluational hierarchy. That is, we show for any ordinal α and sentence A: if $\ulcorner A \urcorner^{\mathbb{N}} \in \sigma_\alpha$, then $+\mathbf{TA} \vdash +A$. The base case is trivial, since $\sigma_0 = \emptyset$.

For the limit step, suppose that λ is a limit ordinal and that for all $\alpha < \lambda$ and sentences A with $\ulcorner A \urcorner^{\mathbb{N}} \in \sigma_\alpha$, it is the case that $+\mathbf{TA} \vdash +A$. Let $\ulcorner A \urcorner^{\mathbb{N}} \in \sigma_\lambda$. Because $\sigma_\lambda = \bigcup_{\alpha<\lambda} \sigma_\alpha$ there is some $\alpha < \lambda$ such that $\ulcorner A \urcorner^{\mathbb{N}} \in \sigma_\alpha$. By assumption, $+\mathbf{TA} \vdash +A$.

For the successor step, we assume that for all $\ulcorner B \urcorner^{\mathbb{N}} \in \sigma_\alpha$ it is the case that $+\mathbf{TA} \vdash +B$ and show that if $\ulcorner A \urcorner^{\mathbb{N}}$ is a member of $\sigma_{\alpha+1}$, then $+\mathbf{TA} \vdash +A$. So let $\ulcorner A \urcorner^{\mathbb{N}} \in \sigma_{\alpha+1}$ and assume that $+\mathbf{TA} \nvdash +A$. Let Γ be the deductive closure of $+\mathbf{TA}$ under $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$. Note that if $\Gamma \vdash \ominus\neg A$ by an evidence-preserving proof, then $\Gamma \vdash +A$ which is not the case by assumption. Thus there is no evidence-preserving proof of $\ominus\neg A$ from Γ and hence $\Gamma \cup \{+\neg A\}$ is consistent*. By Lemma 7.2, $\Gamma \cup \{+\neg A\}$ has a maximally consistent* extension $\hat{\Gamma}$. Let τ be $\{u^{\mathbb{N}} \mid +Tu \in \hat{\Gamma}\}$.

By the induction hypothesis, for all $\ulcorner B \urcorner^{\mathbb{N}} \in \sigma_\alpha$, $+B \in \Gamma$ and since Γ is closed under $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$-inference, $+T\ulcorner B \urcorner \in \Gamma$ and so $+T\ulcorner B \urcorner \in \hat{\Gamma}$ that is $\ulcorner B \urcorner^{\mathbb{N}} \in \tau$. Hence $\tau \supseteq \sigma_\alpha$. We now show that for any $\ulcorner B \urcorner^{\mathbb{N}} \in \sigma_\alpha$, $\ulcorner \neg B \urcorner^{\mathbb{N}} \notin \tau$. Let $\ulcorner B \urcorner^{\mathbb{N}}$ be a member of σ_α. By the induction hypothesis $+\mathbf{TA} \vdash +B$. This means that $+\mathbf{TA} \vdash +\neg T\ulcorner \neg B \urcorner$.

$$(+\neg\text{I.})\ \dfrac{(-\text{T-IN})\ \dfrac{(-\neg\text{I.})\ \dfrac{+B}{-\neg B}}{-T\ulcorner \neg B \urcorner}}{+\neg T\ulcorner \neg B \urcorner}$$

Therefore, $+\neg T\ulcorner \neg B \urcorner \in \Gamma \subseteq \hat{\Gamma}$. Since $\hat{\Gamma}$ is consistent*, $+T\ulcorner \neg B \urcorner \notin \hat{\Gamma}$. By definition of τ, this means that $\ulcorner \neg B \urcorner^{\mathbb{N}} \notin \tau$. One can analogously show that for all $\ulcorner \neg B \urcorner^{\mathbb{N}} \in \sigma_\alpha$, $\ulcorner B \urcorner^{\mathbb{N}} \notin \tau$. So τ fulfills the conditions (i) and (ii) of the van Fraassen hierarchy.

Now, by the previous theorem, $\mathbb{N}, \tau \models \neg A$. Thus there is a τ that is a superset of σ_α fulfilling (i) and (ii) with $\mathbb{N}, \tau \models \neg A$. By definition of $\sigma_{\alpha+1}$, this contradicts the assumption that $\ulcorner A \urcorner^{\mathbb{N}} \in \sigma_{\alpha+1}$. □

Cantini's and the maximally consistent hierarchy. We first prove Theorem 6.3.

Theorem 6.3. The following are derivable in $\mathsf{QWBL}^{\omega}_{\mathsf{TM}}$.

- $+T\ulcorner\neg A\urcorner \leftrightarrow \neg T\ulcorner A\urcorner$
- $+T\ulcorner A \wedge B\urcorner \leftrightarrow (T\ulcorner A\urcorner \wedge T\ulcorner B\urcorner)$.

Recall that $\mathsf{QWBL}^{\omega}_{\mathsf{TM}}$ adds to $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$ the following rules.

$$(+\mathrm{MC})\ \frac{+A_1, +A_2, ..., +A_n \ \vdash\ \bot}{+T\ulcorner A_1\urcorner, +T\ulcorner A_2\urcorner, ..., +T\ulcorner A_n\urcorner \ \vdash\ \bot}\ \begin{array}{l}\text{if the inference to } \bot \text{ uses no premises}\\ \text{signed with } \ominus \text{ and no truth rules.}\end{array}$$

$$(-\mathrm{MC})\ \frac{-A_1, -A_2, ..., -A_n \ \vdash\ \bot}{-T\ulcorner A_1\urcorner, -T\ulcorner A_2\urcorner, ..., -T\ulcorner A_n\urcorner \ \vdash\ \bot}\ \begin{array}{l}\text{if the inference to } \bot \text{ uses no premises}\\ \text{signed with } \ominus \text{ and no truth rules.}\end{array}$$

The proof of Theorem 6.3 proceeds as follows.

Proof. Negation: To derive $+T\ulcorner\neg A\urcorner \rightarrow \neg T\ulcorner A\urcorner$, assume for Smileian *reductio** that $+T\ulcorner\neg A\urcorner \wedge T\ulcorner A\urcorner$. Since $+\neg A, +A \vdash \bot$, by (+MC) $+T\ulcorner\neg A\urcorner \wedge T\ulcorner A\urcorner \vdash \bot$, so by Smileian *reductio** $-T\ulcorner\neg A\urcorner \wedge T\ulcorner A\urcorner$, which entails $+T\ulcorner\neg A\urcorner \rightarrow \neg T\ulcorner A\urcorner$. To derive $+\neg T\ulcorner A\urcorner \rightarrow T\ulcorner\neg A\urcorner$, assume for Smileian *reductio** that $+\neg T\ulcorner A\urcorner \wedge \neg T\ulcorner\neg A\urcorner$. Since $-A, -\neg A \vdash \bot$, by (–MC) $-T\ulcorner A\urcorner, -T\ulcorner A\urcorner \vdash \bot$. So $+\neg T\ulcorner A\urcorner \wedge \neg T\ulcorner\neg A\urcorner \vdash \bot$, which entails $+\neg T\ulcorner A\urcorner \rightarrow T\ulcorner\neg A\urcorner$ by Smileian *reductio**

Conjunction: To derive $+T\ulcorner A \wedge B\urcorner \rightarrow T\ulcorner A\urcorner \wedge T\ulcorner B\urcorner$ assume towards a Smileian *reductio** that $+T\ulcorner A\wedge B\urcorner\wedge\neg T\ulcorner A\urcorner$. By the above result for negation, this entails $+T\ulcorner A\wedge B\urcorner \wedge T\ulcorner\neg A\urcorner$, which entails $\bot$ by (+MC), since $+A \wedge B, +\neg A \vdash \bot$. So $+T\ulcorner A \wedge B\urcorner \rightarrow T\ulcorner A\urcorner$ and analogously $+T\ulcorner A \wedge B\urcorner \rightarrow T\ulcorner B\urcorner$. To derive $+T\ulcorner A\urcorner \wedge T\ulcorner B\urcorner \rightarrow T\ulcorner A \wedge B\urcorner$ assume towards a Smileian *reductio** that $+T\ulcorner A\urcorner \wedge T\ulcorner B\urcorner \wedge \neg T\ulcorner A \wedge B\urcorner$. By the above result for negation, this entails $+T\ulcorner A\urcorner \wedge T\ulcorner B\urcorner \wedge T\ulcorner\neg(A \wedge B)\urcorner$, which entails $\bot$ by (+MC), since $+A, +B, +\neg(A \wedge B) \vdash \bot$. □

The proof of Theorem 6.4 is largely analogous to the above proof of Theorem 6.1.

Theorem 6.4. For all $A \in \mathcal{L}_{A_T}$: $\ulcorner A\urcorner^{\mathbb{N}} \in \sigma^{\mathrm{mc}}$ iff $+\mathbf{TA} \vdash^{\mathsf{QWBL}^{\omega}_{\mathsf{TM}}} +A$.

Recall that σ^{mc} is the fixed point of the hierarchy supervaluating over τ with:

(i.) τ is classically consistent and

(ii.) for any A, either $\ulcorner A\urcorner^{\mathbb{N}} \in \tau$ or $\ulcorner\neg A\urcorner^{\mathbb{N}} \in \tau$.

For the right-to-left direction of Theorem 6.4, the classical consistency condition (i) suffices to show that the truth rules preserve membership in σ^{mc} as in the proof of Theorem 7.1. We need to additionally show that the added rules also preserve membership.

- (+MC). We need to show that if $\ulcorner \neg \bigwedge_{1 \leq i \leq n} A_i \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{mc}}$, then also $\ulcorner \neg \bigwedge_{1 \leq i \leq n} T \ulcorner A_i \urcorner \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{mc}}$. So suppose that the former is the case and assume towards a *reductio* that there is a maximally classically consistent $\tau \supseteq \sigma^{\mathrm{mc}}$ such that $\mathbb{N}, \tau \models \bigwedge_{1 \leq i \leq n} T \ulcorner A_i \urcorner$.

  This means that for all $1 \leq i \leq n$, $\ulcorner A_i \urcorner^{\mathbb{N}} \in \tau$. But by assumption, $\ulcorner \neg \bigwedge_{1 \leq i \leq n} A_i \urcorner^{\mathbb{N}}$ is in σ^{mc} and hence also in τ. So τ is not classically consistent.

- (–MC). We need to show that if $\ulcorner \neg \bigwedge_{1 \leq i \leq n} \neg A_i \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{mc}}$, then also $\ulcorner \neg \bigwedge_{1 \leq i \leq n} \neg T \ulcorner A_i \urcorner \urcorner^{\mathbb{N}} \in \sigma^{\mathrm{mc}}$. So suppose that the former is the case and assume towards a *reductio* that there is a maximally classically consistent $\tau \supseteq \sigma^{\mathrm{mc}}$ such that $\mathbb{N}, \tau \models \bigwedge_{1 \leq i \leq n} \neg T \ulcorner A_i \urcorner$.

  This means that that for all $1 \leq i \leq n$, $\ulcorner A_i \urcorner^{\mathbb{N}} \notin \tau$. Since τ is *maximally* classically consistent, for all $1 \leq i \leq n$, $\ulcorner \neg A_i \urcorner^{\mathbb{N}} \in \tau$. But by assumption, $\ulcorner \neg \bigwedge_{1 \leq i \leq n} \neg A_i \urcorner^{\mathbb{N}}$ is in σ^{mc} and hence also in τ. So τ is not consistent.

For the left-to-right direction of Theorem 6.4, it suffices to observe that the τ provided by the model existence result (Theorem 7.3) are maximally classically consistent and so satisfy condition (i) and (ii) in the definition of the maximally consistent hierarchy.

Proof. Consistency. Suppose there is some A such that $\ulcorner A \urcorner^{\mathbb{N}}$ and $\ulcorner \neg A \urcorner^{\mathbb{N}}$ are both members of τ. Then by definition $+T \ulcorner A \urcorner$ and $+T \ulcorner \neg A \urcorner$ are members of Γ. By Theorem 6.3, $\vdash +T \ulcorner \neg A \urcorner \rightarrow \neg T \ulcorner A \urcorner$, so it follows that $+\neg T \ulcorner A \urcorner \in \Gamma$, but as also $+T \ulcorner A \urcorner \in \Gamma$, Γ is inconsistent*. Contradiction, hence there is no such A.

Maximality. Suppose there is some A such that $\ulcorner A \urcorner^{\mathbb{N}}$ and $\ulcorner \neg A \urcorner^{\mathbb{N}}$ are both *not* members of τ. By definition $+T \ulcorner A \urcorner$ and $+T \ulcorner \neg A \urcorner$ are both not members of Γ. By Theorem 6.3, $\vdash +\neg T \ulcorner A \urcorner \rightarrow T \ulcorner \neg A \urcorner$, so it follows that $+\neg T \ulcorner A \urcorner \notin \Gamma$, but as also $+T \ulcorner A \urcorner \notin \Gamma$, Γ is not maximally consistent*. Contradiction, hence there is no such A. □

We now turn to Theorem 6.2.

Theorem 6.2. For all $A \in \mathcal{L}_{A_T}$: $\ulcorner A \urcorner^{\mathbb{N}} \in \sigma^{\mathsf{C}}$ iff $+\mathbf{TA} \vdash^{\mathsf{QWBL}^{\omega}_{\mathsf{TC}}} +A$.

The set σ^C is the fixed point of the hierarchy supervaluating over τ with: if $\ulcorner A \urcorner^{\mathbb{N}} \in \tau$, then $\ulcorner \neg A \urcorner^{\mathbb{N}} \notin \tau$. And $\mathsf{QWBL}^{\omega}_{\mathsf{TC}}$ adds to $\mathsf{QWBL}^{\omega}_{\mathsf{T}}$ the following rule.

$$\text{(C)}\ \frac{+T\ulcorner\neg A\urcorner}{+\neg T\ulcorner A\urcorner}$$

The proof of Theorem 6.2 proceeds like the proof of Theorem 6.4. In the right-to-left argument, Cantini's definition of consistency suffices for the argument and the rule (C) is easily shown to be sound. For the left-to-right direction, we only need to show that the τ provided by Theorem 7.3 is consistent, for which the rule (C) suffices, since it provides the required part of Theorem 6.3 for negation.

General model theory. A *QWBL*$_T$ *model* is a tuple $(\mathcal{M}, t)$ where $\mathcal{M}$ is a set of models of predicate logic whose domains extend $\mathbb{N}$ and t is a function that maps every $M \in \mathcal{M}$ to a set of numbers so that the set of sentences whose Gödel numbers are in $t(M)$ is classically consistent. Then:

- $M, t(M) \models A$ is defined as usual where $t(M)$ fixes the extension of T.
- $\mathcal{M}, t \models +A$ iff for all $M \in \mathcal{M}$, $M, t(M) \models A$; $\mathcal{M}, t \models -A$ iff for all $M \in \mathcal{M}$, $M, t(M) \not\models A$; and $\mathcal{M}, t \models \ominus A$ iff there is a $M \in \mathcal{M}$ with $M, t(M) \not\models A$.

We say that a QWBL_T model is *T-admissible* if for all A: (for all $M \in \mathcal{M}$, $M, t(M) \models A$) iff (for all $M \in \mathcal{M}$, $\ulcorner A \urcorner^{\mathbb{N}} \in t(M)$). QWBL_T is sound and complete for the class of T-admissible QWBL_T models. Soundness follows from the usual argument and T-admissibility.

For completeness let Γ be a consistent set of formulae. Henkin's construction allows one to find *Henkin extensions* of Γ, that is sets $\hat{\Gamma} \supseteq \Gamma$ that are maximally consistent* and contain witnesses (for all A, if $\hat{\Gamma}$ contains $+\exists x.A$, it also contains $+A[c^A/x]$ where c^A is a constant symbol). For every Henkin extension $\hat{\Gamma}$, let $M^{\hat{\Gamma}}$ be the canonical term model for $\{A \mid +A \in \hat{\Gamma}\}$ and let $\tau^{\hat{\Gamma}} = \{u^{\mathbb{N}} \mid +Tu \in \hat{\Gamma}\}$. The proof of Theorem 7.3 can be straightforwardly adapted to show the following.

Theorem 7.4. $+X \in \hat{\Gamma}$ *iff* $M^{\hat{\Gamma}}, \tau^{\hat{\Gamma}} \models X$.

Because $\hat{\Gamma}$ contains witnesses, the ω-Rule is not needed in the quantifier step.

Now, for every maximally consistent* $\Gamma' \supseteq \Gamma$ let $\hat{\Gamma}' \supseteq \Gamma'$ be a Henkin extension and $M^{\hat{\Gamma}'}$ be its term model. Let $\mathcal{M}$ be the set of all these $M^{\hat{\Gamma}'}$ and define $t(M^{\hat{\Gamma}'}) = \tau^{\hat{\Gamma}'}$. It is easy to see that $(\mathcal{M}, t)$ is T-admissible and, using Theorem 7.4, a model of Γ.

Sound and complete model theories for extensions of QWBL_T can be found by specifying admissibility conditions corresponding to the additional rules governing the truth predicate.

References

Beall, Jc. 2009. *Spandrels of Truth*. Oxford: Oxford University Press.

Beisecker, Dave. 2019. "Denial Has Its Consequences: Peirce's Bilateral Semantics." *Transactions of the Charles S. Peirce Society* 55: 361–86.

Bledin, Justin and Tamar Lando. 2018. "Closure and Epistemic Modals." *Philosophy and Phenomenological Research* 97: 3–22.

Boyd, Richard N. 1983. "On the Current Status of the Issue of Scientific Realism." *Erkenntnis* 19: 45–90.

Brandom, Robert. 1994. *Making it Explicit*. Cambridge, MA: Harvard University Press.

Cantini, Andrea. 1990. "A Theory of Formal Truth Arithmetically Equivalent to ID_1." *Journal of Symbolic Logic* 244–59.

Conee, Earl and Richard Feldman. 2004. *Evidentialism: Essays in Epistemology*. Oxford: Clarendon Press.

Dickie, Imogen. 2010. "Negation, Anti-realism, and the Denial Defence." *Philosophical Studies* 150: 161–85.

Drucker, Daniel. 2020. "The Attitudes We Can Have." *Philosophical Review* 129: 591–642.

Dummett, Michael. 1978. *Truth and Other Enigmas*. Cambridge, MA: Harvard University Press.

———. 1991. *The Logical Basis of Metaphysics*. Cambridge, MA: Harvard University Press.

Dutilh Novaes, Catarina. 2015. “A Dialogical, Multi-Agent Account of the Normativity of Logic.” *Dialectica* 69: 587–609.

Feferman, Solomon. 1991. “Reflecting on Incompleteness.” *The Journal of Symbolic Logic* 56: 1–49.

Field, Hartry. 2008. *Saving Truth From Paradox*. New York: Oxford University Press.

van Fraassen, Bas. 1971. *Formal Semantics and Logic*. New York: Macmillan.

Frege, Gottlob. 1919. “Die Verneinung: Eine logische Untersuchung (Negation: A logical Investigation).” *Beiträge zur Philosophie des deutschen Idealismus* 1: 143–57.

French, Rohan. 2016. “Structural Reflexivity and the Paradoxes of Self-reference.” *Ergo* 3: 113–31.

Gupta, Anil. 1993. “A Critique of Deflationism.” *Philosophical Topics* 21: 57–81.

Halbach, Volker. 2011. *Axiomatic Theories of Truth*. Cambridge: Cambridge University Press.

Halbach, Volker and Leon Horsten. 2005. “The Deflationists’ Axioms for Truth.” In *Deflationism and Paradox*, eds. Jc Beall and Bradley Armour-Garb, 203–17, Oxford: Oxford University Press.

———. 2006. “Axiomatizing Kripke’s Theory of Truth.” *The Journal of Symbolic Logic* 71: 677–712.

Harman, Gilbert. 1986. *Change in View*. Cambridge, MA: MIT Press.

Heck, Richard Kimberly. 2012. “A Liar Paradox.” *Thought* 1: 36–40, originally published under the name “Richard G. Heck, Jr”.

Hlobil, Ulf. 2019. “Faithfulness for Naive Validity.” *Synthese* 196: 4759–74.

Horsten, Leon. 2009. “Levity.” *Mind* 118: 555–81.

———. 2011. *The Tarskian Turn. Deflationism and Axiomatic Truth*. Cambridge, MA: MIT Press.

Horwich, Paul. 1998. *Truth*. Oxford: Clarendon Press.

———. 2010. *Truth – Meaning – Reality*. Oxford: Oxford University Press.

Incurvati, Luca and Julian J Schlöder. 2017. “Weak Rejection.” *Australasian Journal of Philosophy* 95: 741–60.

———. 2021. "Inferential Expressivism and the Negation Problem." In *Oxford Studies in Metaethics*, vol. 16, ed. Russ Shafer-Landau, 80–107, Oxford: Oxford University Press.

———. 2022. "Epistemic Multilateral Logic." *Review of Symbolic Logic* 15: 505–36.

———. forthcoming. *Reasoning With Attitude*. New York: Oxford University Press.

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

Künne, Wolfgang. 2003. *Conceptions of Truth*. Oxford: Oxford University Press.

Lewis, David. 1979. "Scorekeeping in a Language Game." *Journal of Philosophical Logic* 8: 339–59.

McGee, Vann. 1992. "Maximal Consistent Sets of Instances of Tarski's Schema (T)." *Journal of Philosophical Logic* 21: 235–41.

Meadows, Toby. 2013. "Truth, Dependence and Supervaluation: Living with the Ghost." *Journal of Philosophical Logic* 42: 221–40.

———. 2015. "Infinitary Tableau for Semantic Truth." *Review of Symbolic Logic* 8: 207–35.

Misak, Cheryl. 2016. *Cambridge Pragmatism: From Peirce and James to Ramsey and Wittgenstein*. Oxford: Oxford University Press.

Oms, Sergi. 2020. "Minimalism, Supervaluations and Fixed Points." *Synthese* 197: 139–53.

Parsons, Terence. 1984. "Assertion, Denial, and the Liar Paradox." *Journal of Philosophical Logic* 137–52.

Peirce, Charles Sanders. 1905. "What Pragmatism Is." *The Monist* 15: 161–81.

Picollo, Lavinia and Thomas Schindler. 2018. "Deflationism and the Function of Truth." *Philosophical Perspectives* 32: 326–51.

Prawitz, Dag. 2015. "Explaining Deductive Inference." In *Dag Prawitz on Proofs and Meaning*, ed. Heinrich Wansing, 65–100, Dordrecht: Springer.

Price, Huw. 1990. "Why 'not'?" *Mind* 99: 221–38.

Priest, Graham. 1979. "The Logic of Paradox." *Journal of Philosophical logic* 8: 219–41.

Quine, W. V. O. 1970. *Philosophy of Logic*. Cambridge, MA: Harvard University Press.

Ramsey, Frank Plumpton. 1927. "Facts and Propositions." *Aristotelian Society Supplementary Volume* 7: 153–70.

Restall, Greg. 2005. "Multiple Conclusions." In *Logic, Methodology and Philosophy of Science: Proceedings of the Twelfth International Congress*, eds. Petr Hájek, Luis Valdés-Villanueva, and Dag Westerståhl, 189–205, London: King's College Publications.

Richard, Mark. 2008. *When Truth Gives Out*. Oxford: Oxford University Press.

Ripley, David. 2013a. "Paradoxes and Failures of Cut." *Australasian Journal of Philosophy* 91: 139–64.

———. 2013b. "Revising up: Strengthening Classical Logic in the Face of Paradox." *Philosophers Imprint* 13: 1–13.

Rosenblatt, Lucas. 2021. "Bilateralism and Invalidities." *Inquiry* 64: 481–510.

Rumfitt, Ian. 2000. "'Yes' and 'No'." *Mind* 109: 781–823.

Scharp, Kevin. 2013. *Replacing truth*. Oxford: Oxford University Press.

Schindler, Thomas. 2020. "A Note on Horwich's Notion of Grounding." *Synthese* 197: 2029–38.

Schroeder, Mark. 2008. *Being For*. Oxford: Oxford University Press.

Schulz, Moritz. 2010. "Epistemic Modals and Informational Consequence." *Synthese* 174: 385–95.

Sellars, Wilfrid. 1969. "Language as Thought and as Communication." *Philosophy and Phenomenological Research* 29: 506–27.

Smiley, Timothy. 1996. "Rejection." *Analysis* 56: 1–9.

Soames, Scott. 1999. *Understanding Truth*. Oxford: Oxford University Press.

Stern, Johannes. 2018. "Supervaluation-Style Truth Without Supervaluations." *Journal of Philosophical Logic* 47: 817–50.

Tennant, Neil. 1999. "Negation, Absurdity and Contrariety." In *What Is Negation?*, eds. Dov Gabbay and Heinrich Wansing, 199–222, Dordrecht: Kluwer.

Weir, Alan. 1996. "Ultramaximalist Minimalism!" *Analysis* 56: 10–22.

———. 2005. "Naive Truth and Sophisticated Logic." In *Deflationism and Paradox*, eds. Jc Beall and Bradley Armour-Garb, 218–49, Oxford: Oxford University Press.

———. 2015. “A Robust Non-Transitive Logic.” *Topoi* 34: 99–107.

Williams, Michael. 1988. “Epistemological Realism and the Basis of Scepticism.” *Mind* 97: 415–39.

Williamson, Timothy. 2000. *Knowledge and Its Limits*. Oxford: Oxford University Press.

Wright, Crispin. 1992. *Truth and Objectivity*. Cambridge, MA: Harvard University Press.

Yalcin, Seth. 2007. “Epistemic Modals.” *Mind* 116: 983–1026.