FIX, EXPRESS, QUANTIFY.
DISQUOTATION AFTER ITS LOGIC.

CARLO NICOLAI

1. Deflationism

Truth-theoretic deflationism is a view about the nature and function of the truth *predicate*. I will characterise it by means of four main theses:

(FIX) the meaning of 'is true' is fixed by the Tarski-biconditionals " 'A' is true if and only if A".

FIX can be traced back to Frege (1918) and (Quine, 1970, §1). In its propositional version, it is also present in Ramsey (1927). Horwich (1998) and Field (1994) are (fairly) recent, primary references.

(EXPRESS) the purpose of truth is to express – in virtue of FIX – infinite conjunctions and disjunctions.

(QUANTIFY) truth is *fundamentally* a device to perform sentential quantification over pronominal variables.

Forcefully proposed by (Quine, 1990, §33); a formal rendering of Quine's claim has been put forth in Halbach (1999) – see also Heck (2005) for a discussion.

(EXPLAIN) Truth does not play a substantial role in philosophical and scientific explanations.

REMARK 1. In a recent rendering due to Shapiro (1998) and Ketland (1999), the thesis that the theory of truth should not interfere with the non-semantic world by entailing new, non-semantic information: in a slogan, truth should be *conservative* (Cieśliński, 2017). I will not consider EXPLAIN in what follows – but I have discussed it in Nicolai (2015, 2016).

The main aim of the paper is to establish that the former key theses of deflationism are incompatible with the fundamental properties of reasonable, primitive truth predicates. In particular, I will argue:

- that FIX leads to the adoption of dialetheism, and that disquotationalism shouldn't be tied to such nonclassical option;
- that the combination of EXPRESS and QUANTIFY leads to the claim that an infinite conjunction and the assertion 'all conjuncts are true' should be equivalent in a strong sense. But they cannot be.
- that even if one considers QUANTIFY in isolation, the claim that the truth predicate fulfils the theoretical role of higher-order quantification in a first-order setting is highly dubious.

2. Fix

FIX is first and foremost a thesis about the *meaning* of 'is true'. I start by focusing on the logical structure of its core principle:

$$\mathtt{Tr}\ulcorner A\urcorner \leftrightarrow A \tag{T}$$

with A a sentence of a sufficiently expressive language, possibly containing $\mathtt{Tr}$ itself. Of course the $\mathtt{Tr}$-schema cannot possibly be right in such form, due to the Liar paradox. Horwich (1998) proposes to consider only non-problematic instances of it. This isn't a trivial task (McGee, 1992).

It suffices here to restrict myself to instances of (T) featuring an even number of negation symbols. I call (PT) the resulting schemata

$$\mathtt{Tr}\ulcorner A\urcorner \leftrightarrow A, \qquad\qquad \mathtt{F}\ulcorner \bar{A}\urcorner \leftrightarrow A, \tag{PT}$$

for all A of $\mathcal{L}^+$ – i.e. the positive fragment of $\mathcal{L}_{\mathtt{Tr}} = \mathcal{L}_{\mathbb{N}} \cup \{\mathtt{Tr}\}$, with $\bar{A}$ the dual of A.

CLAIM 1. (PT) *cannot be taken to express simple material equivalence: A correct reading of the equivalence should at least assign to it a certain modal status.*

One reason for this is that (PT) *nor any consistent subset of* (T) can deliver in full generality basic laws of truth of the form:[1]

$$\forall\varphi,\psi \in \mathcal{L}_{\mathbb{N}}\,(\mathtt{Tr}(\varphi\wedge\psi) \leftrightarrow \mathtt{Tr}\varphi \wedge \mathtt{Tr}\psi). \tag{$\wedge$}$$

Bristol Philosophy Seminar, May 2, 2019.

[1]Notational conventions: I employ $A, B, C, \ldots$ to range over concrete, standard sentences of the language; by contrast $\varphi, \psi, \chi, \ldots$ range over 'nominalised', quantifiable sentences.

or even quasi-logical trivialities of the form

(1) $$\forall \varphi, \psi \in \mathcal{L}_{\mathbb{N}} \, (\mathtt{Tr}\,\varphi \to \mathtt{Tr}\,(\varphi \vee \psi)).$$

Another reason is that such claim is a core component of classical deflationism. (Field, 1994, §6): cognitive equivalence. (Horwich, 1998, §3, f. 5),(Field, 1994): interaction with alethic modalities. Fischer et al. (2019): justificatory role. Halbach (2003): truth-analyticity.

The 'minimal' modal principles – notice that $\Box$ is a predicate – that I adopt amount to a *generalised* version of Halbach's 2003 theory:

(2) $\forall \varphi \in \mathcal{L}^{\Box}_{\mathtt{Tr}} (\mathtt{Prov}_{\mathtt{fol}}(\varphi) \to \Box\varphi)$ (i.e. classical logical truths are boxed);

(3) $\forall \varphi, \psi \in \mathcal{L}^{\Box}_{\mathtt{Tr}} (\Box(\varphi \to \psi) \wedge \Box\varphi \to \Box\psi)$ (i.e. closure under classical conditional);

(4) $\forall \varphi \in \mathcal{L}^{+} \Box((\mathtt{Tr}\,\dot{\varphi} \leftrightarrow \varphi) \wedge (\mathtt{F}\,\dot{\varphi} \leftrightarrow \varphi))$ (i.e. all instances of the disquotation schemata are boxed)

(5) if A is a consequence of the theory in $\mathcal{L}_{\mathtt{Tr}}$, then also $\Box A$ is;

(6) $\forall \varphi \in \mathcal{L}_{\mathtt{Tr}} (\Box\varphi \to \mathtt{Tr}\,\varphi^{*})$ – where $\cdot^{*}\colon \mathcal{L}_{\mathtt{Tr}} \to \mathcal{L}^{+}$ is such that $(\neg \mathtt{Tr}\,x)^{*} = \mathtt{F}x^{*}$.

The new principles indeed overcome the expressive limitations of material disquotation, for instance they entail the *modal version* – and therefore also the non-modal version – of ($\wedge$):

$$(\forall \varphi, \psi \in \mathcal{L}^{+}) \, \Box\big(\mathtt{Tr}\,(\varphi \mathbin{\dot{\wedge}} \psi) \leftrightarrow \mathtt{Tr}\,\dot{\varphi} \wedge \mathtt{Tr}\,\dot{\psi}\big).$$

However, if one considers a liar sentence $l = \ulcorner \neg \mathtt{Tr}\, l \urcorner$, one obtains that $\mathtt{Tr}\, l \wedge \mathtt{F}\, l$:

Proposition 1. *Any weak syntax theory plus* (2)-(6) *has an inconsistent truth predicate.*

Compare this to:

> . . . a central tenet of the point of view advanced here is that the theory of truth and the theory of logic have nothing to do with one another. (Horwich, 1998, §24)

Of course a dialetheist may be happy with prop. 1: I am not prepared to accept inconsistent notions so easily (Fischer et al., 2019).

3. Express, Quantify

3.1. **Infinite sets and generalisations.** The links between express and quantify are clear if one looks at canonical examples. The infinite conjunction

(7) $$\text{(snow is white} \to \text{snow is white)} \wedge \text{(grass is green} \to \text{grass is green)} \wedge \ \ldots$$

is equivalent, via fix, to

(8) $$(\mathtt{Tr}\ulcorner\text{snow is white} \to \text{snow is white}\urcorner) \wedge (\mathtt{Tr}\ulcorner\text{grass is green} \to \text{grass is green}\urcorner) \wedge \ldots$$

Being sentences in quotations terms standing for names of sentences, the truth predicate enables one to affirm

(9) all sentences of the form $\sigma \to \sigma$ are true.

The disquotationalist's claim is then that (9) *expresses* (7), via (8). How should we understand this relation?

Option 1. Definability in a structure: compare sets definable in $\mathcal{L}_{\mathtt{Tr}}$ and the ones definable in certain infinitary languages. Crucially relies on *semantic truth*, so against the disquotationalist tenets.

Option 2. Conservativeness (Halbach, 1999): relative to a minimal theory of truth bearers S, one requires S plus the infinite conjunction of the elements of a set of sentences R to yield *the same non-semantic consequences* as S+'all elements of R are true'.

Problem 1: this is false if one considers both infinite conjunctions and disjunctions (Heck, 2005).

Problem 2: Only $\mathtt{Tr}\ulcorner A \urcorner \to A$ is sufficient for the claim. There is nothing special about disquotation that enables one to express (7) by means of (9).

As forcefully argued already by Gupta (1993), there are independent reasons to require something much stronger, more akin to *sameness of meaning* between (7) and (9):

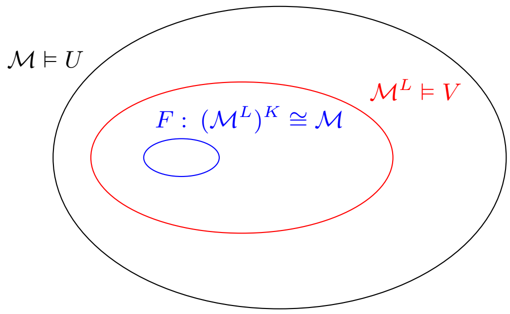
(i) The way in which fix is standardly presented refers to an *analysis* of 'is true' (Field, 1994; Horwich, 1998), not a simple necessary equivalence.

(ii) This is also in accordance with certain *intended applications* (Horwich, 1998, pp. 44-47): law-like generalisations such as 'true beliefs tend to facilitate success' should have the same explanatory role as the infinite conjunction of sentences of the form: 'Subject S wants X; S believes that by doing Y she will achieve X'. Necessary equivalence does not suffice in general for equivalence of explanatory status. An analogous point can be made for truth-functional laws (Field, 1994, pp. 258-9).

To study this equivalence between infinite conjunctions and general claims that preserves explanatory/theoretical status, I find it natural to resort to standard notions of *theoretical equivalence*.

Let U, V be first-order theories extending a basic syntax theory (equivalently, finite set theory, basic arithmetic):

Definition 1 (Retract). *U is a* retract *of V iff there are interpretations $K\colon U \to V$ and $L\colon V \to U$ and there is a U-definable isomorphism F between $L \circ K$ and* id_U.



Definition 2. BI-INTERPRETABILITY *U and V are* bi-interpretable *if there are relative interpretations $K\colon U \to V$ and $L\colon V \to U$ s.t. the following both hold:*

(i) *U is a retract of V, witnessed by $F, L \circ K, \mathsf{id}_U$.*

(ii) *V is a retract of U, witnessed by $G, K \circ L, \mathsf{id}_V$.*

REMARK 2. Under natural assumptions, bi-interpretability entails definitional equivalence (Quine, 1975; Glymour, 1970; Barrett and Halvorson, 2016).

Back to our original comparison between (7) and (9):

PROPOSITION 2. *Suppose a set of sentences S of $\mathcal{L}_B$ is represented in our syntax base B via the formula $P(x)$. Suppose further that S (i.e. $F(\cdot)$) is not finitely axiomatisable over B. Let B^{Tr} an extension of B with a finite set of principles entailing at least all typed instances of the $\mathtt{Tr}$-schema. The theories*

- $B + \{P(\ulcorner A \urcorner) \to A \mid A \in S\}$
- $B^{Tr} + \forall \varphi (P(\varphi) \to \mathtt{Tr}(\varphi))$

are not *bi-interpretable – let alone definitionally equivalent.*

Proof Idea. The proof rests on the fact that bi-interpretability preserves finite axiomatisability over B: therefore, if U and V were bi-interpretable, we would be able to find a finite axiomatisation of S over B, which contradicts our assumption. □

I take Proposition 2 to be strong evidence against the claim that *any* form of strong reduction of infinite sets of sentences to truth-theoretic generalisations *can ever succeed*.

3.2. **Truth as a quantifier.** The discussion above highlights the quantificational role of truth. A reaction may be to understand it as a second-order quantifier (Picollo and Schindler, 2017): i.e. to forget about EXPRESS and focus on QUANTIFY. The tools introduced above tell us that *this fails even at the most simple level.*

Let us start with (relational version of) the language of second-order arithmetic $\mathcal{L}_2$ and consider the basic arithmetical theory PA extended on the one hand with *arithmetical, parameter-free predicative comprehension* ($\mathsf{ACA}^{\mathsf{pf}}$):

($\mathsf{CA}^{\mathsf{pf}}$) $\quad \exists X \forall u (u \in X \leftrightarrow A(u))$ $\quad$ in A no second-order variables,

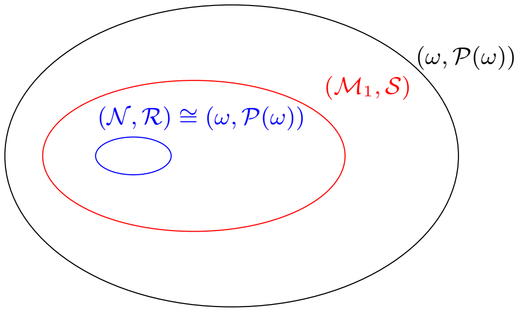
on the other by *typed uniform disquotation* (UTB):

(UTB) $\quad \forall x\, (\mathtt{Tr}\ulcorner A(\dot{x}) \urcorner \leftrightarrow A(x))$ $\quad$ for all $A(v)$ the purely arithmetical language $\mathcal{L}_{\mathbb{N}}$.

Induction *is extended* to truth and second-order variables.

PROPOSITION 3 (Nicolai (2017)). $\mathsf{ACA}^{\mathsf{pf}}$ *and* UTB *are* not *bi-interpretable.*

That UTB is a retract of $\mathsf{ACA}^{\mathsf{pf}}$ follows from the standard interpretation. The converse, however, cannot hold:



The conceptual point is that the truth predicate is a 'bad quantifier', disallowing the richness of interpretations of the 'properties' involved.

4. Conclusion

Recent developments of truth-theoretic deflationism have focused on a formal analysis of principles of truth and their logical properties. In this extended sense, any position that considers truth as a primitive concept, and characterizes it *only* by means of a simple set of axioms – or rules of inference –, would count as deflationary (Horsten, 2012). I have attempted to reconcile these formal approaches to the classical tenets of truth-theoretic deflationism: FIX, EXPRESS, QUANTIFY, and EXPLAIN.

The upshot of the analysis seems clear:

- The best approach available to make explicit the modal status of disquotation in FIX leads to inconsistency. And this is likely to generalize to structurally similar accounts of FIX.
- The combination of EXPRESS and QUANTIFY requires that infinite lots of sentences and their corresponding truth theoretic generalizations stand in a close theoretical relationship that preserves their explanatory status. In the most natural way of understanding this relationship, that is via formal notions of conceptual or theoretical equivalence, such relationship simply cannot exist.
- Finally, even if one considers QUANTIFY in isolation, by claiming that the principles of the deflationist's truth predicate are there *just* to mimic higher-order quantification in a first-order setting, one requires the theoretical equivalence between truth and quantification. In the natural sense of definitional equivalence or biinterpretability, this equivalence cannot hold.

What has been said, of course, does not impact on truth theoretic deflationism intended in a loose sense, as the formal and philosophical study of principles of truth. If, however, the deflationary approach to truth is charaterised – as it is commonly done – by FIX, EXPRESS, QUANTIFY, and EXPLAIN, its chances of success are slim.

References

Barrett, T. W. and Halvorson, H. (2016). Glymour and Quine on theoretical equivalence. *Journal of Philosophical Logic*, 45(5):467–483.
Cieśliński, C. (2017). *The Epistemic Lightness of Truth: Deflationism and its Logic*. Cambridge University Press.
Field, H. (1994). Deflationist views of meaning and content. *Mind*, 103(411):249–285.
Fischer, M., Horsten, L., and Nicolai, C. (2019). Hypatia's silence. Truth, justification, and entitlement. *Noûs*. Forthcoming.
Frege, G. (1918). Thoughts. In Frege, G., editor, *Logical Investigations*. Blackwell.
Glymour, C. (1970). Theoretical realism and theoretical equivalence. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1970:275–288.
Gupta, A. (1993). A critique of deflationism. *Philosophical Topics*, 21(1):57–81.
Halbach, V. (1999). Disquotationalism and infinite conjunctions. *Mind*, 108(429):1–22.
Halbach, V. (2003). Modalized disquotationalism. In Horsten, L. and Halbach, V., editors, *Principles of Truth*, pages 75–102. De Gruyter.
Heck, R. (2005). Truth and disquotation. *Synthese*, 142(3):317–352.
Horsten, L. (2012). *The Tarskian Turn*. MIT University Press, Oxford.
Horwich, P. (1998). *Truth*. Clarendon Press.
Ketland, J. (1999). Deflationism and tarski's paradise. *Mind*, 108(429):69–94.
McGee, V. (1992). Maximal consistent sets of instances of tarski's schema (t). *Journal of Philosophical Logic*, 21(3):235–241.
Nicolai, C. (2015). Deflationary truth and the ontology of expressions. *Synthese*, 192(12):4031–4055.
Nicolai, C. (2016). A note on typed truth and consistency assertions. *Journal of Philosophical Logic*, 45(1):89–119.
Nicolai, C. (2017). Equivalences for truth predicates. *Review of Symbolic Logic*, 10(2):322–356.
Picollo, L. and Schindler, T. (2017). Disquotation and infinite conjunctions. *Erkenntnis*, 83:899–928.
Quine, W. V. (1970). *Philosophy of Logic*. Harvard University Press.
Quine, W. V. (1975). On empirically equivalent systems of the world. *Erkenntnis*, 9(3):313–28.
Quine, W. V. (1990). *Pursuit of Truth*. Harvard University Press.
Ramsey, F. P. (1927). Facts and propositions. *Proceedings of the Aristotelian Society*, 7(1):153–170.
Shapiro, S. (1998). Proof and truth. *Journal of Philosophy*, 95(10):493–521.