Annali del Dipartimento di Filosofia (Nuova Serie), XIII (2007), pp. 117-145
ISSN 1824-3770 (online) © 2008 Firenze University Press

On some recently debated issues in the theory of formal truth

Riccardo Bruni

As the title suggests, this paper aims at surveying some recent advances in the theory of formal truth. It contains an account of the debate concerning the deflationist approach to truth, according to which truth is a 'thin' notion in that it should involve no assumption of whatsoever nature. We review here the main issues that were comprised by the discussion accompanying the attempts of translating this idea into logical terms. In the second half of the paper, we focus on a recent theory of truth proposed by Hartry Field, a former 'champion' of the deflationary approach. We then discuss it both with respect to the previous conceptual account, and to some further observation concerning the truth–as–revision machinery that this theory can be proved to implicitly make use of.

Keywords: *Tarski's theory of truth, deflationism, revision theory of truth, quasi–inductive definitions.*

1. *Introduction: logical power vs. metaphysical bareness*

Part of the formal investigation on the notion of 'truth' as it has been pursued in recent years, goes back to Tarski's seminal contribuition to the field in a peculiar way. It originates from regarding it as providing controversial indications. Roughly speaking, this contrast is to be found in Tarski's solution to the semantical paradoxes of the natural language as based on (i) a definition of truth that satisfies some very basic requirements, and (ii) on distinguishing between the object language, to which truth refers, and its metalanguage, where truth values of sentences are calculated on the basis of inductive clauses.

Condition (i) is condensed in the well known T(arski's)-convention

$$\text{“}p\text{” is true if and only if } p$$

whereas (ii) is made necessary in order to make it immune to the effects of antinomies.

The point is that the T-convention, once it is given the value of a *definition*, seems to justify the idea that truth is basically an insubstantial notion in the sense that it conveys no new content to the sentence itself (at least no content which requires a specific ontological committment), while Tarski's additional construction leads to conclude that truth is a powerful, higher-order notion (as confirmed by the theorem on the undefinability of truth in formal languages which are adequate for arithmetic). Philosophically speaking: in the first sense, truth is metaphysically 'thin', while in the second one it is committed to compelling ontological assumptions.

It appeared thus coherent to those who accepted the T-convention as a definition of truth, to refute Tarski's distinction. Their efforts were spent in the direction of making this standpoint more precise, and logically coherent.

Hartry Field was among the champions of this view, and significantly took part in the enterprise of providing support for it.[1] His latest contributions,[2] are centered on a peculiar investigation over a possible assessment of the unrestricted Tarski's schema by making use of some non-classical interpretations of the logical connectives. By doing that, as we shall argue subsequently, he seemingly departed from the position he had previously supported (or, more simply, he provided an alternative

[1] See H. Field, *Deflationist Views of Meaning and Content*, «Mind», 103, 1994, pp. 247–285, and, by the same author, *Deflating the Conservativeness Argument*, «The Journal of Philosophy», 96, 1999, pp. 533–540.

[2] H. Field, *The Semantic Paradoxes and the Paradoxes of Vagueness*, in J. C. Beall, *Liars and Heaps*, Oxford University Press, 2003, pp. 262–311, H. Field, *A revenge–immune solution to the semantic paradoxes*, «Journal of Philosophical Logic», 32, 2003, pp. 139–177.

to it). Also, some results on Field's latest proposal allow us to look at it as a truth definition which is equivalent, in a sense to be defined, to processes falling under the belief–revision approach, and thus originated by motivations which are different from Field's. In turn, this makes it possible to suggest how to possibly deepen the research in this direction, by re–considering Field's construction in this wider context.

This last comment notwithstanding, the purpose of the present paper is largely expository, the basic aim being that of surveying some of the most recent literature on certain conceptual aspects involved in the above–mentioned approach to truth, as well as technical aspects of Field's latest proposal. In the first direction, we have chosen to stick to the attempts of proof–theoretically assessing the standpoint in question as a mean for discussing it on a more rigorous basis.

2. *The conceptual framework: the deflationary approach*

Proof-theoretically speaking, by the *deflationary* (equivalently, *disquotationalist* or *minimal*) theory of truth it is usually meant, at least since Horwich's book on truth,[3] the standpoint which regards everything that must be assumed about truth to be comprised into all 'safe' instances (namely, all those instances which do not yield paradoxical conclusions) of the T(arski's)–schema

$$\mathrm{T}(\overline{\phi}) \leftrightarrow \phi \tag{1}$$

where $\overline{\phi}$ is some 'name' for the sentence ϕ. In absence of a general criterion for establishing which instances of the above should be regarded as 'safe',[4] we can take the definition we have just given to refer to the schema (1) instanciated by all ϕ's in which the truth predicate T does not occur.

[3] P. Horwich, *Truth*, Basil Blackwell, 1990.

[4] See the results presented in V. McGee, *Maximal Consistent Sets of Instances of Tarski's Schema (T)*, «Journal of Philosophical Logic», 21, 1992, pp. 235–241.

From a more conceptual point of view, the deflationary approach to truth is usually described as the idea that truth is characterized by:

(i): its being an 'insubstantial' notion;

(ii): its allowing *fertile generalizations* (which it would be impossible to make if truth is dispensed with).

Condition (i) reflects the idea, which is partially encompassed by Tarski's schema, that truth must entail only consequences which are, say, *purely semantical* in character. This is sometimes referred to by saying that truth must be a 'thin notion', ontologically speaking: it matters only for those 'facts' where truth itself or related notions are somewhat implicitly involved. As we suggested, this aspect is directly recognizable in the 'moral' of the T-schema: truth does not add anything new to what the sentence itself already says.

However, this assumption must not be taken to be equivalent to the idea that this is *always* the case. On the contrary, condition (ii) is quite explicit in stressing that truth allows for augmenting our knowledge significantly, since the absence of it makes it *impossible* to state some specific statements. Deflationists have particularly emphasized two cases in which we face examples of this sort: (a) in the case of *blind abscriptions*, that is when we committ ourselves to the truth of a sentence *indirectly*, i.e. without mentioning it explicitly (as in the phrase, e.g., «The sentence written on page ... of the book " ... " is true.»); (b) in the case of the concise expression of an infinite amount of sentences (as in «All the axioms of the (first order) system of Peano Arithmetic are true»).[5]

In the first case, the elimination of truth via the T–schema is impossible since we do not know the sentence it applies to. In the second, it is equally impossible because doing that would require an infinite amount of resources.

[5] The second of the two remarks, belongs to a view with notable tradition which goes back at least to Quine (see W. V. O. Quine, *Philosophy of Logic*, Harvard University Press, 1970, p. 12, and, by the same author, *Pursuit of Truth*, Harvard University Press, 1990, p.81).

A minor point, which is somewhat implicit in the above committment of the deflationary theory of truth to the T–schema, is that deflationists refuse the usual assessment of (1) as done by Tarski's himself, i.e. by distinguishing between language and meta-language. As a consequence, truth for a language $\mathcal{L}$ is entirely comprised in a language $\mathcal{L} \cup \{T\}$, with the (necessary) addition of a *caveat* on those instances of (1) which should be allowed.

Due to some critical accounts and relative replies, these features of the standpoint we are concerned with, which were originally formulated quite informally and unprecisely, have undergone a more serious investigation in recent years, and gained a more definite shape by that.

Let us take for example the claim that truth must be insubstantial. It has been stressed[6] that this claim necessarily binds deflationists to accept the idea that a proper deflationary theory of truth should be conservative over every other theory truth axioms are added to (and then, over logic itself – which, in turn, would allow to obtain conservativity over every finitely axiomatized theory by a simple application of the deduction theorem). Shapiro's idea is then that if conservativeness did not hold, then insubstantiality would fail.

Fixing some terminology may help understanding where this may turn out to represent a relevant issue to the deflationary approach. To stick to the case which will particularly interest us in the following, given a system of axioms S written in a given base language, and a theory T, which is based on the language obtained by the given one by adding a fresh unary predicate T for truth, and which further contains the axioms for it, then we say that T *is conservative over* S if and only if $\mathsf{S} \vdash \phi$ whenever $\mathsf{T} \vdash \phi$, for every formula ϕ of the language of S. Thus, that T is conservative over S can really be taken as entailing that truth,

[6] Particularly, by Stewart Shapiro in his *Proof and Truth: Trough Thick and Thin*, «The Journal of Philosophy», 95, 1998, pp. 493–521, and *Deflation and Conservation*, in V. Halbach and L. Horsten, *Principles of Truth*, Hansel–Hohenhausen, 2002, pp. 103–128.

as characterized by the axioms added to the given system, is insubstantial in the precise sense that it does not allow to prove anything new with respect to the original formal framework.

This part of Shapiro's standpoint was reinforced by the sympathetic attitude toward it which was expressed by proper deflationists, either by directly making this claim on their own, or by indirectly trying to circumvent those problems which arise in the attempt of attaining to that requirement, and which are taken by Shapiro as a disproof of the deflationary approach.[7]

Shapiro's criticism in this latter respect is comprised in a general argument which is based on Gödel's second incompleteness theorems. Recently, this very same criticism has been re–stated (read: reinforced, once acceptance of the conservativeness argument is conceded) in a much more simple form by Volker Halbach:[8] the pure logic of identity suffices to prove that the theory obtained by adding the unparadoxical instances of the T-schema can't be conservative over logic.[9]

Conservativity over logic however, might be a deflationary line of response, really seems to be too much. This could be motivated not just by the way in which the formal investigation on truth is *really* carried out, namely by adding the axioms for truth to a given *base theory*, but it can also be argued for on the basis of the kind of insubstantiality deflationists themselves seem to have in mind, i.e. as something referring on a pre–existing ontology.

[7] See Hartry Field, *Deflating the Conservativeness Argument*, cit.

[8] V. Halbach, *How Innocent is Deflationism*, «Synthese», 126, 2001, p. 179.

[9] The argument goes as follows: by the fact that $\forall x(x = x)$ is trivially provable by the logic of identity, and $\forall x(x \neq x)$ is trivially refutable by it, it follows that in the theory augmented by Tarski's schema, $T(\overline{\forall x(x = x)})$ and $\neg T(\overline{\forall x(x \neq x)})$ are derivable. But then it suffices to make use of the Leibniz's principle

$$x = y \rightarrow P(x) \leftrightarrow P(y)$$

to derive both $\overline{\forall x(x = x)} \neq \overline{\forall x(x \neq x)}$, and $\exists x \exists y(x \neq y)$, and thus to conclude that truth implies the existence of at least two distinguished object (and thus that it cannot be neither conservative over logic, nor metaphysically neutral).

Nevertheless, passing from conservativeness over logic to conservativeness over a base theory doesn't seem to be of help.[10]

Existing proof–theoretical results are focused on two theories in particular: one, call it $\mathsf{Tr(PA)}$, which is obtained from PA by the addition of the axioms for the inductive closure of the truth predicate, and where the induction schema is extended to all formulas of the language $\mathcal{L}^+ = \mathcal{L}_{\mathsf{PA}}$ *plus* the truth predicate T; the other, $\mathsf{Tr(PA)} \restriction$, which is defined as the latter system except for the fact that the induction schema is restricted to all formulas of the base language $\mathcal{L}_{\mathsf{PA}}$ only.[11]

The results in question go as follows: $\mathsf{Tr(PA)}$ is significantly stronger than the base theory (it has the same arithmetical theorems of the second order theory ACA where the comprehension schema is restricted to arithmetical formulas[12]), while $\mathsf{Tr(PA)} \restriction$ is conservative over PA. Thus, only the second one, if any, might be of help in the attempt of circumventing the conservativeness critique.

However, and this is an important point to stress, these two theories entail a complete change in the attitude toward truth. Truth axioms, in fact, are not coped with by means of instances of the T–schema, but make use of the inductive definition of truth in a formal language.[13] Before making any use of the

[10] This is the point emphasized by authors like L. Horsten and J. Ketland (see L. Horsten, *The Semantical Paradoxes, the Neutrality of Truth and the Neutrality of the Minimalist Theory of Truth*, in P. Cortois, *The Many Problems of Realism*, Tilburg University Press, 1995, pp. 173–187, and J. Ketland, *Deflationism and Tarski's Paradise*, «Mind», 108, 1999, pp. 69–94.

[11] See S. Feferman, *Reflecting on Incompleteness*, «Journal of Symbolic Logic», 56, 1991, pp. 1–49 and A. Cantini, *Logical Frameworks for Truth and Abstraction. An Axiomatic Study*, Amsterdam: Elsevier, 1996. For a very broad historical survey of this and related issues, see also A. Cantini, *Paradoxes, Self-Reference and Truth in the 20th Century*, in D. M. Gabbay and J. Woods, *Handbook of the History of Logic, vol. 5: Logic from Russell to Church*, Elsevier, 2008 (announced date).

[12] See, e.g., S. G. Simpson, *Subsystem of Second Order Arithmetic*, Springer, 1999 for details about this system.

[13] Informally speaking, these axioms establish closure of the truth predicate by induction on the formulas of the language. Thus, there would be an axiom stating, for any k-ary predicate P^k and any sequence $t_1, \ldots, t_k$ of closed terms,

results above, it is then a priority for a deflationist to provide some justification for the conceptual shift involved herein.

Before going into that, however, it's useful to go on surveying also the present–day debate on the second feature which, according to what we said above, characterizes the deflationary approach, namely the idea that truth should serve for making generalizations which would be impossible without it. This idea has been made more precise by Halbach,[14] insofar as it concerns the possibility of expressing infinite conjunctions by means of the truth predicate.

Let A be a *definable* set of sentences in the base language $\mathcal{L}$, that is a given starting language *without* the truth predicate. This means that for some formula ϕ of $\mathcal{L}$, and a model $\mathbf{M}$ of $\mathcal{L}$,[15] we have that $A = \{a \mid \mathbf{M} \models \phi(\overline{a})\}$ (where, for the sake of simplicity, we may assume that the base language contains all the names $\overline{a}$ for any object a of the universe of $\mathbf{M}$). Then, it's easy to check that expressing the infinite conjunction of all the elements of A (which corresponds to the utterance of all of them), coincides with the truth in $\langle \mathbf{M}, T \rangle$, where $T \subseteq \mid \mathbf{M} \mid$ provides the interpretation of the truth predicate in the model, of the sentence:

$$\forall x(\phi(x) \to T(x)) \tag{2}$$

in the extended language.

The first result which can be proved on this is a negative one:[16] for any formula $\phi(x)$ the extension of which (that is,

that $P(t_1, \ldots, t_k)$ is true if and only if P^k holds for the values of the $t_1, \ldots, t_k$ (for a given internal definition of a function evaluating terms, i.e. assigning them to their code). Furthermore, there would be unsurprising axioms for the truth of compound formulas (for the exact list of principles, see e.g. S. Feferman, *Reflecting on Incompleteness*, cit., pp. 13–14). By an induction argument, these axioms can be proved to entail all required instances of Tarski's schema.

[14] See V. Halbach, *Disquotationalism and Infinite Conjunctions*, «Mind», 108, 1999, pp. 1–22.

[15] For the sake of the argument, this model should be 'standard' in the sense of fullfilling the requirements for it to be *acceptable* in the sense of Y. Moshovakis, *Elementary Induction on Abstract Structures*, North Holland, Studies in Logic 77, 1974.

[16] See V. Halbach, *Disquotationalism and Infinite Conjunctions*, cit., p. 10.

the collection $\{n \mid \mathbf{M} \models \phi(\overline{n})\}$) is infinite in any given model $\mathbf{M}$ of a base theory S, $\forall x(\phi(x) \rightarrow T(x))$ is not a theorem of S + T-schema.[17]

In order to let the original deflationary idea survive, one might first observe that (2) is by no means the only way to express the conjunction of infinitely many formulas, and that the *schema* $\phi(\overline{\psi}) \rightarrow \psi$ already does the job.[18] Conceptually speaking, the rationale behind this sort of reply would be that, according to the deflationary approach, it is not important for the sentence representing infinite conjunctions to logically imply the *truth*, or even the *provability* of every conjunct, but *only* to logically imply every conjunct.

Then, the following is shown to hold:[19] (i) $\forall x(\phi(x) \rightarrow T(x))$ obviously implies $\phi(\overline{\psi}) \rightarrow \psi$ via Tarski's schema, and (ii) the theories $\mathsf{S}+\{\phi(\overline{\psi}) \rightarrow \psi \mid \psi \text{ sentence}\}$ and $\mathsf{S}+\{T(\overline{\psi}), \forall x(\phi(x) \rightarrow T(x))\}$ have the same consequences over the formulas of the base language of S (for any formulas ϕ and ψ – the latter not containing T). Result (i) entails also that the truth predicate is indispensable in order to express infinite conjunctions as comprised by the schema $\phi(\overline{\psi}) \rightarrow \psi$, since this (that is, (i)) holds for arbitrary ϕ while there are known instances of that schema which are known not to be expressible by a single sentence or a finite set of sentences otherwise.[20]

In the light of all the results we've quoted, the literature offers two main solutions to the controversy we have dealt with in the present section:

[17] For the formulation of this result to be precise, one should refer to the extension of a formula as containing those n which are codes of sentences. This assumption was skipped here for the sake of readability, on the basis of those tricks which allow to expand a given coding of a language onto a proper subset of an infinite set, into a mapping to the infinite set as whole by allowing infinite repetitions (i.e., by dropping the injectivity of the code function).

[18] The verification that for $A = \{a \mid \mathbf{M} \models \phi(\overline{a})\}$, $\mathbf{M} \models \phi(\overline{\psi}) \rightarrow \psi$ if and only if, for all $\psi \in A$, $\mathbf{M} \models \psi$ is a simple exercise.

[19] V. Halbach, *Disquotationalism and Infinite Conjunctions*, cit., pp. 13–14.

[20] This is the case, for example, even of the more simple version of the so-called 'reflection principles', i.e. statements expressing validity for a certain system of axioms S in the language of the system itself.

Solution 1 [Field 1999]: the minimal approach to truth can escape criticisms *á la* Shapiro, by embracing the inductive axiomatization of truth in $\mathsf{Tr}(\mathsf{PA}) \restriction$ in substitution to safe instances of the T–schema. This move can be justified on the basis of the *pure truth theoretic nature* of the inductive clauses (in the sense that, for example, they are independent from the concept of number they make use of via arithmetization), as it can be appreciated by confronting them with the additional number–theoretic axioms of the theory in question. Notice that since the inductive clauses are provably stronger than instances of the T–schema, they serve, as the latter does, for the expression of infinite conjunctions (Halbach's argument above). In this sense, deflationary truth preserves both its generalization power and, via the conservativeness result on $\mathsf{Tr}(\mathsf{PA}) \restriction$, its 'innocence'.[21]

Solution 2 [Halbach 1999–2001]: a deflationist should circumvent Shapiro's criticism *ab ovo*, namely by denying any committment to conservativeness over logic, which is provably impossible. If the purpose of his account of truth is just generalizations and conservativeness over a base theory, then the original idea of sticking to the safe instances of the T–schema already suffices.[22] This would put him in a far better position than arguing in favor of the 'pure nature' of some assumption can ever do, since it is unclear how a 'purely truth–theoretic axiom' should read. However, the best move whatsoever would be to accept the idea that allowing generalizations which were

[21] We are perhaps forcing Field's intention a little bit here by making of him a supporter of this line of response. A more honest statement would be to say that he seemingly sympathized with it (see Hartry Field, *Deflating the Conservativeness Argument*, cit.).

[22] This statement relies on the theorem on generalizations we mentioned above, and on a result on (syntactical) conservativeness which is contained in V. Halbach, *Conservative Theories of Classical Truth*, «Studia Logica», 62, 1999, Lemma 2.1. Halbach's viewpoint has been partly criticized very recently by Gary Kemp in his *Disquotationalism and Expressiveness*, «Journal of Philosophical Logic», 34, 2005, pp. 327–332, according to whom the deflationary claim about the T–schema as allowing generalizations in the sense of Halbach's result should be rejected as a complete failure.

contrarily unavailable, is indeed a *strong* requirement which is likely to imply the provability of new 'substantial' statements over a given theory. Thus, deflationists should focus more decisively on the fertile side of truth, and drop the dogma of innocence.

These conceptual replies to criticisms of the deflationary approach notwithstanding, Field has very recently offered new contributions which seem to suggest a third solution to the above debate. Their feature is to go back reinforcedly to the original idea of starting from the (unrestricted) Tarski's schema, in a peculiar way to be made precise below. As we shall argue in our final comments, the most remarkable aspect of these works is that they suggest a way in which the logical investigation along these lines can be further deepened.

3. *Saving full Tarski's schema: Field's latest proposal*

The account of the previous section shows that the verification of the main pillars of deflationism by means of the tools of proof theory, provides uncertain results. It was maybe for related reasons, although there is no explicit claim on this, that in most recent times Hartry Field, who, as we have seen, actively took part into the above debate, focused on an attempt of justyfing what we may refer to as a *minimal version* of deflationism. As we shall explain in a moment, the choice of this name is not entirely satisfactory, and it must be thus taken *cum grano salis*.

The basic idea is to provide new grounds for accepting a naive theory of truth, that is one which keeps Tarski's schema *in its unrestricted formulation* (i.e. with no constraints on the formulas to which truth applies, thus allowing in particular self–application), without being forced to accept any sort of language/metalanguage distinction.

So, while in fact this aim reduces the deflationary assumptions to a minimum (dropping in particular any consideration concerning the specific purposes truth is committed to attaining to), it comprises nonetheless a radical formulation of the original prescription which is retained. This strengthening, when combined with the language–vs.–metalanguage refutation, necessarily entails, for the sake of consistency, that the implication connective with respect to which Tarski's schema is formulated, must fail to satisfy one of the logical properties which are known to hold for the classical implication.[23]

Field's general idea is to show how to expand a classical model $\mathbf{M}$ for a given language $\mathcal{L}$ based on the usual alphabet of first order predicate logic (plus some possible extra features), to a non-classical model $\mathbf{M}^*$ for the language $\mathcal{L}^{++}$, which expands $\mathcal{L}$ by means of a new implication connective *and* a truth predicate.

Since this construction requires that the starting theory S based on $\mathcal{L}$ be adequate for arithmetic, we assume, for our expository purposes, that our given theory contains Peano Arithmetic PA.[24] Further, we assume S to be classical, logically speaking.

Thus, the language $\mathcal{L}$ is based on a standard alphabet for first order logic where $\neg, \vee, \forall$ are taken as primitive logical symbols, and $\wedge, \supset, \equiv, \exists$ are defined by the usual clauses,[25] plus special symbols $0, succ, +, \times, =$ with the usual meaning. Formulas and terms of $\mathcal{L}$ are inductively generated as usual. The axioms of S are those of a complete formulation of classical first-order logic plus the usual principles of PA, and some possible further principle.

[23] Our reference to a *logic* here can be misleading, for, as we shall see later, Field's construction is basically (and, in a sense to be made precise, *essentially*) semantical. This proviso may apply also to further notions to be defined subsequently.

[24] To be more precise, Field's definition assumes the given theory to allow to explicitly define a predicate $N(x)$ satisfied in a standard model of arithmetic by all and only the natural numbers, and a similar definition of the usual arithmetical operations. Furthermore, the theory must provide an adequate theory of finite sequences in order to perform the arithmetization of syntax. Thus, our assumption could be further weakened by taking Robinson's arithmetic as a base theory, or the theory usually called $\Sigma^0_1 - IA$.

[25] We follow here Field's suggestion to use Russell's symbol $\supset$ for the classical implication as defined by $(A \supset B) := (\neg A \vee B)$ (and $A \equiv B$ for $(A \supset B) \wedge (B \supset A)$), while we retain the arrow $\rightarrow$ for the non-classical one.

By $\mathcal{L}^{++}$ we indicate instead the language which is obtained by adding an additional connective ($\rightarrow$) and the truth predicate T to $\mathcal{L}$. Let $\mathbf{M}$ be a model of S which is assumed to be standard for PA, and let us indicate by $\mid \mathbf{M} \mid$ its domain. We assume each formula A of $\mathcal{L}^{++}$ be given a Gödel code $\overline{A}$ by means of some standard Gödel numbering. Then, for any ordinal numbers α, β, with $\beta \leq \Gamma$ (Γ representing the next greater cardinal than the one of the cardinality of $\mathbf{M}$[26]), Field shows how to construct a sequence of three-valued models $\langle \mathbf{M}_{\alpha,\beta} \rangle$ for sentences (closed formulas) of $\mathcal{L}^{++}$ (thus assigning to any of such a sentence a truth value among $\{0, \frac{1}{2}, 1\}$), where, for any $\alpha, \beta \in ON$, we have $\mid \mathbf{M}_{\alpha,\beta} \mid = \mid \mathbf{M} \mid$.

This sequence is constructed in a lexicographical order (namely, building $\mathbf{M}_{\alpha,\beta}$ prior to $\mathbf{M}_{\alpha',\beta'}$ iff $\alpha < \alpha'$ or $\alpha = \alpha'$ and $\beta < \beta'$), according to inductive satisfaction clauses entailing the following conditions:

- atomic formulas of the original language $\mathcal{L}$ retain the value they have with respect to the model $\mathbf{M}$;
- keeping α fixed, formulas of the form $T(t)$ (where t is a certain closed term of $\mathcal{L}^{++}$) are assigned a value along the sequence $< \mathbf{M}_{\alpha,\beta} >_{\beta \in ON}$ with $\beta \leq \Gamma$ according to a Kripkean construction: thus, for each $\beta \leq \Gamma$, the interpretation $T_{\alpha,\beta}$ of T in $\mathbf{M}_{\alpha,\beta}$ is given by the collection of the (codes of) sentences which are true (i.e., get value 1) in some $\mathbf{M}_{\alpha,\beta'}$ with $\beta' < \beta$ (where its complement, which provide the interpretation for $\neg T(t)$, is given by the collection of all non–codes of sentences *plus* codes of sentences previously evaluated as false ones); $\frac{1}{2}$ is then assigned to terms falling in the remaining cases.

[26] This means, e.g, that $\Gamma = \aleph_1$ in case $\mathsf{S} \equiv \mathsf{PA}$. The reason for the restriction to ordinals less or equal to Γ as second element of the couples $< \alpha, \beta >$ is due to the existence of fixed points, of which Γ is an obvious example, once the monotonicity lemma is given (see below).

- the inductive step involving the usual logical connectives $\neg, \wedge, \forall$, is provided by the strong-Kleene satisfaction clauses;[27]
- finally, formulas of the form $A \to B$ are evaluated in any $\mathbf{M}_{\alpha,\beta}$ of the sequence above according to the following procedure: get value 1 those formulas of this form whose antecedents have the latest of 'Kripkean' values (i.e., the value in $\mathbf{M}_{\delta,\Gamma}$ with Γ being the ordinal explained above) which is less or equal to the one of the consequents *all along an interval* $[\gamma, \alpha)$ (including the first and excluding the second), for some $\gamma < \alpha$; get value 0 formulas such that for all such an interval the latest value of A is strictly greater than the value of B; and $\frac{1}{2}$ otherwise.[28]

Field's main results on this construction comprise, first of all, a reformulation of Kripke's monotonicity and fixed point theorems, which, in this case, read as follows:[29]

Monotonicity: For every α ordinal and every sentence A, if $\|A\|_{\alpha,\beta} \in \{0,1\}$ then $\|A\|_{\alpha,\gamma} = \|A\|_{\alpha,\beta}$ for every $\beta \leq \gamma$.

Fixed Point theorem: For every α there exists a $\beta(\alpha) < \Gamma$ such that

$$\|A\|_{\alpha,\beta(\alpha)} = \|A\|_{\alpha,\gamma}$$

for every $\gamma \geq \beta(\alpha)$.[30]

[27] Thus: $\|\neg A\|_{\alpha,\beta} := 1 - \|A\|_{\alpha,\beta}$, $\|A \wedge B\|_{\alpha,\beta} := min\{\|A\|_{\alpha,\beta}, \|B\|_{\alpha,\beta}\}$ and $\|\forall x A\|_{\alpha,\beta} := min\{\|A[x := t]\|_{\alpha,\beta} \mid t \text{ closed term}\}$.

[28] In formal terms, the conditions for truth and falsity of a sentence of the form $A \to B$ read:

$$\|A \to B\|_{\alpha,\beta} = 1 \quad \Leftrightarrow \quad \exists\gamma\forall\delta < \alpha(\gamma \leq \delta \to \|A\|_{\delta,\beta} \leq \|B\|_{\delta,\beta})$$
$$\|A \to B\|_{\alpha,\beta} = 0 \quad \Leftrightarrow \quad \exists\gamma\forall\delta < \alpha(\gamma \leq \delta \to \|A\|_{\delta,\beta} > \|B\|_{\delta,\beta})$$

This corresponds to take the (boolean) truth values of a sentence of that form to be the *inferior limit* of its truth values in the interval $[\gamma, \alpha)$ interpreting → in a classical fashion. This allows to view → as a transfinite iterate of the classical $\supset$ connective, with the liminf operation taking care of the limit levels.

[29] H. Field, *A revenge–immune solution to the semantic paradoxes*, cit., p. 143.

[30] On the basis of this result we may substitute $\beta(\gamma)$ for Γ in the clauses of the previous definition. The very proof of this results explains why in our sequence of models we can confine ourselves to those $\mathbf{M}_{\alpha,\beta}$ where $\beta \leq \Gamma$. Indeed, given

The fixed point theorem itself yields as a consequence the main results on Tarski's schema: by an easy argument,[31] it indeed implies that, for every formula A and every α, $\|A\|_{\alpha,\beta(\alpha)} = \|T(\overline{A})\|_{\alpha,\beta(\alpha)}$. In turn, from this it immediately follows by the →–evaluation clauses, that, for every A and every α

$$\|A \leftrightarrow T(\overline{A})\|_{\alpha,\beta(\alpha)} = 1$$

Furthermore, the same result entails the intersubstitutivity of A with $T(\overline{A})$.[32]

Some further properties which is worth mentioning, are obtained as follows. Let us define for any A the *ultimate value* of A ($[[A]]$) as follows:[33]

$$[[A]] := \begin{cases} 1, \text{ iff } \exists\alpha\forall\gamma(\alpha \leq \gamma \rightarrow \|A\|_{\gamma} = 1) \\ 0, \text{ iff } \exists\alpha\forall\gamma(\alpha \leq \gamma\, \|A\|_{\gamma} = 0) \\ \frac{1}{2}, \text{ otherwise} \end{cases}$$

The main reason for introducing this definition is what Field calls the Continuity Lemma, which establishes that values of $\|A \rightarrow B\|_{\alpha}$ are continuous at limit ordinals.[34]

Then, Field's main theorem states that there exist *acceptable points* for the ultimate value of any A, namely:

the monotonicity result, an easy cardinality argument on the set of sentences of $\mathcal{L}^{++}$ makes it clear that there must be some ordinal $\delta \leq \Gamma$ such that $\|A\|_{\alpha,\delta} = \|A\|_{\alpha,\delta+1}$.

[31] See H. Field, *A revenge–immune solution to the semantic paradoxes*, cit., p. 143.

[32] Where, as expected, this means that, for every formula B and every α, $\|B\|_{\alpha,\beta(\alpha)} = \|B[A := T(\overline{A})]\|_{\alpha,\beta(\alpha)}$.

[33] After the above fixed–point result, it is clear that the truth value that matters is the one the formula gets at that level. Following Field, we omit to mention explicitly the fixed points and thus we simply write $\|A\|_{\alpha}$ for $\|A\|_{\alpha,\beta(\alpha)}$, for every sentence A and every ordinal α.

[34] That is, for any limit ordinal λ we have:

$$\|A \rightarrow B\|_{\lambda} := \begin{cases} 1, \text{ iff } \exists\alpha < \lambda\forall\beta \in [\alpha,\lambda), \|A \rightarrow B\|_{\beta} = 1 \\ 0, \text{ iff } \exists\alpha < \lambda\forall\beta \in [\alpha,\lambda), \|A \rightarrow B\|_{\beta} = 0 \\ \frac{1}{2}, \text{ otherwise} \end{cases}$$

By previous considerations it easily follows that $\big[[T(\overline{A}) \leftrightarrow A]\big] = 1$.

Main theorem: For every sentence A there exists an ordinal Δ such that $[[A]] = \|A\|_\Delta$.

As to the complexity of this Δ, Field's argument gives no indication at all. Later in his paper,[35] he proves that Δ is *not* less than a certain recursive ordinal λ_0 which is used in a transfinite iteration of a *determinately true* operator on formulas, that we will have the occasion to mention in the subsequent section.

As we shall see in §3.2, it turns out that this estimate is still too vague (and defective), and that its corrected form may be used as part of some criticisms to be raised against Field's proposal as we shall do in the final remarks.

3.1. *Some properties of Field's construction*

According to what we have said so far, it should be clear that the mechanism of Field's definition can be informally described as generally being of the one involved in the so–called «revision theory»of truth,[36] with a Kripkean component (something which can be made more precise on the basis of some recent results that we will mention in the following). The latter is responsible for the evaluation of the T–part of the language $\mathcal{L}^{++}$ (provided an evaluation of the $\rightarrow$–part of it be given), while the evaluation of the $\rightarrow$–part, as well as its combinations with the T–part, is accounted for in a revision–like fashion. This turns out clearly by a brief description of what happens at the very first levels of the model construction.

By definition, formulas of the language $\mathcal{L}$, that do not involve T and $\rightarrow$, retain the value they take in $\mathbf{M}$ (0 or 1, since $\mathbf{M}$ is classical by assumption) at level $< 0, 0 >$, and they keep this at all subsequent levels since the construction is clearly semantically conservative in this sense over $\mathcal{L}$. Field's machinery is thus specifically conceived to account for the part of $\mathcal{L}^{++}$ which is not in $\mathcal{L}$.

[35] See H. Field, *A revenge–immune solution to the semantic paradoxes*, cit., p. 162.

[36] See A. Gupta and N. Belnap, *The Revision Theory of Truth*, MIT Press, 1993.

Formulas of the form $A \to B$ get value $\frac{1}{2}$ at level $< 0, 0 >$ since the $\{0, 1\}$–evaluation clauses, which refer to the previously constructed levels, fail vacuously. The Kripkean steps which follow (i.e., those from $< 0, 0 >$ to $< 0, \beta(0) >$, or $< 0, \Gamma >$ equivalently) only concern formulas of the form $T(t)$, or which have these sort of formulas as immediate subformulas. Once again we must distinguish between $\mathcal{L}$ and $\mathcal{L}^{++}$: in case T is applied to (codes of) formulas of the former (or to terms which are not codes of sentences), the evaluation of the T–formulas and their logical combinations yields their ultimate value already at $< 0, \beta(0) >$ (even though, for each α as a first component their value is always re–calculated by definition). All other formulas get a value which can be either 'provisional', or the ultimate one according to further assignement of values to the involved $\to$–subformulas.

Once the procedure beneath Field's construction is made clear in this way, and once the results for the Kripkean component of it are given, it is natural to ask how general is the scope of these results. This applies in particular to the fixed point theorem.

It is easily seen that there are, so to say, permanently 'flipping' formulas (i.e., formulas A such that, for every α, a $\beta > \alpha$ can be found for which $\|A\|_\alpha \neq \|A\|_\beta$).[37] A particularly significant example of such a sentence, is provided by the so–called 'Curry sentence', that is the formula K such that for any α, $\|K\|_\alpha = \left\|(T(\overline{K}) \to \bot)\right\|_\alpha$. The semantical behaviour of such a sentence can be easily determined by an easy direct calculation.[38]

[37] Note that the existence of these sentences does not conflict with the Main Theorem on acceptable points we've mentioned in the previous sections: the ultimate value of flipping formulas is in fact simply $\frac{1}{2}$ by definition, even though this is not their 'real' truth value.

[38] By the corollary on intersubstitutivity we have $\|K\|_\alpha = \|K \to \bot\|_\alpha$. But: $\|K \to \bot\|_0 = \frac{1}{2}$ by definition, and hence $\|K\|_0 = \frac{1}{2}$ which in turn implies $\|K \to \bot\|_1 = 0 = \|K\|_1$. So, $\|K \to \bot\|_2 = 1 = \|K\|_1$, and so on for any finite ordinal, which implies $\|K \to \bot\|_\omega = \frac{1}{2} = \|K\|_\omega$. The result referred to in the main text, is obtained by a straightforward generalization.

This yields the following result:

$$\|K\|_\alpha := \begin{cases} \frac{1}{2}, \text{ if } \alpha \text{ is 0 or a limit} \\ 0, \text{ if } \alpha \text{ is an odd successor ordinal} \\ 1, \text{ if } \alpha \text{ is an even successor} \end{cases}$$

Obviously, 'processing' sentences of this sort at acceptable points is very simple, considered that by definition a sentence such as K above has $\frac{1}{2}$ as its ultimate value. But this is just a matter of definition, and it does not seem to allow the introduction of anything like Kripke's distinction between *grounded* and *ungrounded* sentences.[39] Nor, it would serve any purpose to try to make use of his (Kripke's) definition of a *paradoxical* sentence as the one for which there exists no fixed point at which it gets a value, for clearly the absence of *global* fixed points would cause it to apply vacuously.

This suggests that a criticisms to Field's construction might be directed to its inability of explaining on what basis sentences which are neither true or false, featuring those which, like K, the Liar, or the Truth Teller (which are easier to treat, since they can be formulated in the $\rightarrow$-free fragment of the language[40]), are recognized as being intuitively paradoxical, can be said to be, say, 'defective'.

In order to circumvent this, Field introduces an iterable formula operator $D(A) := (\top \rightarrow A) \wedge A$, which seems to provide a reasonable answer in some relevant case.[41] Instead of going

[39] See S. Kripke, *Outline of a Theory of Truth*, «The Journal of Philosophy», 72, 1975, p. 706.

[40] See H. Field, *A revenge–immune solution to the semantic paradoxes*, cit., §5.

[41] It solves, for example, the case of the Liar sentence (see H. Field, *A revenge–immune solution to the semantic paradoxes*, cit., p. 157–161). The whole proposal is intended as answering to a 'revenge'–argument against Field's construction (for the explanation of which we refer the reader to Field's own). On the scope and correctness of this revenge–immune claim by Field for his theory, a recent work by P. Welch and A. Rayo (*Field On Revenge*, in J. C. Beall, *The Revenge of the Liar*, Oxford University Press, to appear (announced date: february, 2008), pp. 617–706), has shown that it depends on assumptions concerning the expressive power of the language under discussion (it may fail, in particular, with respect to languages which are capable of expressing higher–order notions).

into details about this route, however, we will try to retain of Field's discussion of it only those aspects which help deepening an enquire concerning the *logical properties* underlying this construction. This will better serve the purposes of the present contribution.

In this sense, we can make use of the information contained in the central part of Field's paper, which is devoted to a semantical investigation.[42] In particular, Field provides a partial axiomatization[43] of the logic of $\rightarrow$, according to which this connective proves to have some nice properties featuring identity, conjunction and disjunction laws (included distributivity), and one direction of contraposition as valid axioms, whereas modus ponens (which can be generalized as a cut–rule), monotonicity, and a-fortiori appear among the valid rules.

However, as we said, the Fieldian implication is not classical. Among the rules that fail in the case of $\rightarrow$, we have the importation and permutation of the premises, and, moreover, contraction. Thus, it becomes natural to enquire about the relationship between the two sorts of implication, $\supset$ and $\rightarrow$. As done by Field, this investigation leads the excluded middle to play a crucial role.

As a matter of fact, assume that we write $\Phi \vdash_F A$, $\Phi \vdash_F^{TND} A$ for: «$\Phi \models_F A$ is shown to hold by possibly making use of axioms and rules belonging to Field's partial axiomatization of $\rightarrow$», and «$\Phi \models_F A$ is shown to hold as before, *plus* by making extra use of instances of the law of exluded middle involving formulas occurring in Φ itself», respectively. Then, it can be shown, for

[42] The following definition of *logical consequence* for any sentence A of $\mathcal{L}^{++}$, and any set Φ of sentences of the same language is assumed: we say that A is consequence of Φ with respect to Field's construction (and we write it as $\Phi \models_F A$), if and only if $\|\Phi\| = 1 \Rightarrow \|A\| = 1$, with $\|\Phi\| = \inf\{\|B\| \mid B \in \Phi\}$. The corresponding notion of *validity* for a sentence A of $\mathcal{L}^{++}$, is obtained from the latter definition by taking $\Phi = \emptyset$.

[43] This means that only soundness of these axioms and rules with respect to Field's construction can be proved. The problem of completeness turned out to be a serious one (see below).

any formula of $\mathcal{L}^{++}$, that:

$$\begin{array}{rll} A \supset B & \vdash_F & A \rightarrow B \\ & \vdash_F^{TND} & (A \supset B) \rightarrow (A \rightarrow B) \\ (A \rightarrow B) & \vdash_F^{TND} & A \supset B \\ & \vdash_F^{TND} & (A \rightarrow B) \rightarrow (A \supset B) \end{array}$$

Quite naturally, this result is further generalized in a corollary according to which full classical reasoning is allowed in case appropriate instances of the excluded middle are added to Field's axiomatization of the non–classical implication.[44]

The immediate implication of these results can be stated as follow: *full classical reasoning*, featuring the treatement of → as the ordinary implication, *is allowed everywhere the excluded middle holds*. This turns out to be relevant for an approach to truth theories as superstructures built upon some pre–existent domain (which *is* actually the case for Field's construction, as we have defined it). For, the statement we have just stated makes it possible for a classical treatement of the Fieldian implication to be 'imposed from the outside' by adding TND as some sort of non–logical assumption, and letting it to hold on whatever sentence, or, more generally, domain, it seems appropriate (read: safe, or justified) to do.

[44] More precisely, Field obtains a theorem stating that:

$$\Phi \models A \Rightarrow \Phi^* \cup \{B \vee \neg B \mid B \sqsubseteq_a \Phi^* \cup \{A^*\}\} \vdash_F A^*$$

where Φ, A belong to the →–free fragment of $\mathcal{L}^{++}$, $\models$ is the classical relation of logical consequence, Φ^*, A^* are obtained from the given one, substituting → for each occurrence of ⊃, and $\sqsubseteq_a$ is the relation of 'atomic subformula' (see H. Field, *A revenge–immune solution to the semantic paradoxes*, cit., p. 153–154).

3.2. *The complexity issue*

Some additional investigations on Field's proposal has been recently done by Philip Welch.[45] A quick look at Welch's results will make the whole picture more exhaustive, and our subsequent discussion of it a bit more understandable.

Welch succeded indeed in giving an exact answer to the question *How far is it reasonable to iterate the process of calculating truth values according to Field's rules?*, which was left open in Field's paper. It turns out in fact that the least acceptable ordinal is the ordinal ζ, which is the least ordinal admitting a proper Σ_2–extension in the universe of Gödel's constructible sets.[46]

This theorem moves incomparably upward Field's original lower bound result (which we briefly mentioned at the end of §3), since this ζ is not only non–recursive, but far beyond the first non–recursive ordinal ω_1^{CK}.[47]

In addition, three further observations come out as a byproduct of Welch's proof strategy.

First, it turns out that the collection of all Fildian stable truths is (1–1) recursively isomorphic to the set of all the Σ_2–formulas which are true at L_ζ. Similarly, this very same class is isomorphic to the collection of the formulas which are stably true under Herzberger revision procedure (which is known to be recursivly isomorphic to the L_ζ Σ_2 truths as well[48]), and which

[45] See P. Welch, *On Gupta–Belnap Revision Theories of Truth, Kripkean Fixed Points and the Next Stable Set*, «The Bulletin of Symbolic Logic», 7, 2001, pp. 345–360, and, by the same author, *On Revision Operators*, «Journal of Symbolic Logic», 68, 2003, pp. 689–711.

[46] This means that ζ is the least ordinal for which there exists an ordinal $\rho > \zeta$ such that $L_\zeta \prec_{\Sigma_2} L_\rho$, that is such that any Σ_2–formula in the language of set theory is true at the level L_ζ of Gödel's hierarchy of constructible sets if and only if it is true at level L_ρ.

[47] Both ζ and ω_1^{CK} belong to the collection ADM^* of admissible ordinal and their limits. If $\langle \tau_\iota \mid \iota < \omega_1 \rangle$ is an enumeration of such a collection, then it results that $\tau_0 = \omega, \tau_1 = \omega_1^{CK}$ and $\tau_\zeta = \zeta$.

[48] See J. P. Burgess, *The Truth is Never Simple*, «Journal of Symbolic Logic», 51, 1986, Theor. 13.1, 14.1.

is obtained by iterating the Tarskian jump over the empty set and closing under the liminf operation at limits.[49]

Welch's result on the recursive isomorphism has a second fact as an immediate consequence: Field's 'logic' for →, which is presented as an instance of the Logic of Circular Concepts in the spirit of Gupta and Belnap book on the revision theory of truth,[50] *is not axiomatisable.*

As a third, and final, consequence, the complexity measurement can be turned into a proof–theoretical indication: it turns out in fact that in order to perform the calculation of the semantical value of the formulas of $\mathcal{L}^{++}$ one needs a powerful subsystem of second–order number theory (one properly extending $\Delta^1_3 - CA$[51]).

4. *Concluding remarks*

The result having the most immediate significance among those we have quoted in the previous section, is the recursive isomorphism between the set of stable truths under Field's construction and the same set as provided by Herzberger's sequence, which definitely gives a precise shape to the revision–like behaviour of Field's proposal. In turn, this makes the results about complexity less surprising: revision theories of truth are in fact *known to be based on complex procedures.*

[49] That is, the so–called 'Herzberger sequence over the null hypothesis' $\langle H_\alpha \rangle_{\alpha \in On}$ is defined by the clauses $H_0 := \emptyset$, $H_{\alpha+1} := j_T(H_\alpha)$, and $H_\lambda := \liminf_{\beta \to \lambda} H_\beta = \bigcup_{\alpha<\lambda} \bigcap_{\alpha<\beta<\lambda} H_\beta$, for λ limit ordinal (where $j_T(H_\alpha)$ contains the codes of all the formulas of a chosen starting language *plus* the truth predicate T, which are true at H_α according to the usual clauses for Tarskian truth – the interpretation for T being the set H_α itself). Thus, e.g., if the base language is that of first-order logic plus possible additional constants, H_1 contains the set of all classical first–order tautologies.

[50] A. Gupta and N. Belnap, *The Revision Theory of Truth*, cit..

[51] The theory in question, a subsystem of full second–order number theory where the comprehension axiom is restricted to Δ^1_3-formulas, has in fact a least β–model (a least model which is absolute for well–ordering relations) occurring at a level of the constructible hierarchy built up over the reals which is sensibly smaller than the ζ above (and thus which is does not suffice for the definition of the set of stable truths).

Generally speaking, revision constructions can be described as falling under peculiar variations of the usual schema for inductive definitions, namely transfinite generation procedures of sequences $\langle X_\alpha \mid \alpha \text{ ordinal}\rangle$ of, say, sets of natural numbers such that:

$$\begin{array}{lrcl} & X_\alpha & \subseteq & \mathbb{N},\ \alpha \text{ ordinal} \\ (3) & X_{\alpha+1} & := & d(X_\alpha) \\ & X_\lambda & := & G(\langle X_\beta \mid \beta < \lambda\rangle),\ \lambda \text{ limit ordinal} \end{array}$$

(where d and G are operators yielding sets of numbers out of sets of numbers and sequences of sets of numbers, respectively[52]). Notice that here we are not putting any special constraint on d and G as it is usually done, on the contrary, in the most common cases of inductive definitions, and these can be set–theoretic operators whatsoever (in particular, d can be, and it is usually taken to be in all interesting cases, *non-monotonic*, i.e. it *fails* to satisfy $S \subseteq S' \Rightarrow d(S) \subseteq d(S')$).

Since in absence of these constraints the usual definition of inductively defined set of natural number in terms of *closure ordinal* and *fixed points* of the above process is known *not* to apply, the above statement concerning the complexity of revision processes can be made precise in two ways:

(1) Revision processes *as truth definitions*, are complicated: let $\Box\mathrm{T}$ be the collection of *stable truths* of a language $\mathcal{L}^* := \mathcal{L} \cup \{T\}$ over a process $\langle X_\alpha \mid \alpha \text{ ordinal}\rangle$, that is the collection

$$\Box\mathrm{T} := \{\overline{A} \mid A \text{ is (the code of) } \mathcal{L}^*\text{-formula} \wedge \exists\alpha\forall\beta \geq \alpha(\overline{A} \in X_\beta)\}$$

where the operator d here is chosen so to code the Tarskian jump of a set X (see footnote 48), and G is such that (i) $\Box\mathrm{T}_{<\lambda} \subseteq X_\lambda$, and (ii) $X_\lambda \cap (\omega \setminus \Box\mathrm{T}_{<\lambda})$ (with

[52] We're sticking here to just a simple case of this sort of definitions. A more general schema would be to consider Γ as possibly yielding sets of sets of numbers as an output (an example of such a limit operator being the one used in the so–called *Belnap sequences*, where X_λ is identified with the collection of all sets which are *coherent* with the sequence up to λ, i.e. satisfy (i) and (ii) as defined at points 1 and 2 below in the list below in the main text).

$\Box\mathrm{T}_{<\lambda}$ being the restriction of $\Box\mathrm{T}$ to ordinals less than λ). Then $\Box\mathrm{T}$ *forms at least a complete* Π^1_2 *set*;[53]

(2) Revision processes *as set-theoretical definitions*, are complicated as well: take all sets of natural numbers which are *coded by stable elements of revision processes*, namely the collection of those $Z \subseteq \mathbb{N}$ such that:

$$Z = \bigcap\{X^+_{<\infty} \mid \langle X_\alpha \mid \alpha \text{ ordinal}\rangle \text{ rev. proc based on } \ d, G\}$$

where (a) $d : \mathcal{P}(\mathbb{N}) \to \mathcal{P}(\mathbb{N})$ is such that for some formula φ of $\mathcal{L}^* := \mathcal{L} \cup \{P_Y\}$ (where P_Y is a fresh unary predicate coding Y-elementhood), and for every $X \subseteq \mathbb{N}$

$$d(X) = \{n \mid \langle \mathbb{N}, X\rangle \models \varphi(\overline{n})\}$$

(b) $X^+_{<\infty} = \{n \mid \exists\alpha\forall\beta \geq \alpha(n \in X_\beta)\}$, and (c) where G is assumed to satisfy (i) and (ii) of the previous point with $X^+_{<\lambda}$ replacing $\mathrm{T}_{<\lambda}$.

Then this collection *coincide with the* Π^1_2*–definable reals.* Further, the sets of natural numbers which are *strongly coded* by stable elements of revision processes (those $Z \subseteq \mathbb{N}$ such that both Z and $\mathbb{N} \setminus Z$ are coded by stable elements), *are the* Δ^1_2*–definable reals.*[54]

Needless to say, these remarks make it difficult to see in what sense Field's latest proposal could be seen as a proper assessment for a deflationist notion of truth: it is hard in fact to imagine how the deflationist's preoccupation for the metaphysical neutrality of truth could co–exist with the essentially set–theoretical nature of Field's construction, owing to the usual attitude toward set theory which has been historically regarded as a threat for metaphysical issues in the philosophy of mathematics.

[53] See J. P. Burgess, *The Truth is Never Simple*, cit., P. Kremer, *The Gupta–Belnap Systems $S^\#$ and S^* are not axiomatisable*, «Notre Dame Journal of Formal Logic», 34, 1993, pp. 583–596, P. Welch, *On Revision Operators*, cit. A given coding, with no special features, of formulas of the language $\mathcal{L}^*$ is obviously presupposed by this definition.

[54] See P. Welch, *On Gupta–Belnap Revision Theories of Truth, Kripkean Fixed Points and the Next Stable Set*, cit., Theor. 2.1 for a proof.

Of course, it is important to stress that Field never claimed that his construction should be regarded as a proposal to be read in the deflationary line of thought. Moreover, it would be hard to see how such a claim could be consistently made in view of Welch's result concerning the isomorphism between Field's notion of 'ultimate value', and 'stable truth' in Herzberger sequences: as stated by Gupta in particular, the aim of the revision–approach to truth is tailored neither to capture the idea of truth as a primitive notion, nor to support the view that truth is given 'once and for all' (two ideas deflationists seem to be somewhat committed to).

Further, it is also true that the results we have mentioned in the previous section seem to fulfill Field's explicit aim. The result about the theory needed for calculation of the semantical value of formulas, for example, can be related to Field's claim that his construction is Tarskian *only* in the sense that full Tarski's schema remains valid, while no distinction between a language and its own meta–language is provided. Indeed, the truth values for formulas of the language $\mathcal{L}^{++}$, can be calculated in a subsystem of ZFC properly stronger than $\Delta^1_3 - CA$ (thus, to be precise, in a theory based on a *proper sublanguage* of $\mathcal{L}^{++}$ if, as Field does, $\mathcal{L}$ is allowed to contain $\mathcal{L}_{ZFC}$).

However, that very same result also entails the basic difficulties (essential need of higher–level notions, non–axiomatizability) for Field's proposal to be put in the deflationary 'tradition', which, in turn, seem to yield, as a combined moral of the two parts of our survey, a view that makes the deflationist's goal resemble a chimerical search: if the language/metalanguage distinction is banned, metaphysical bareness can only be gained at the cost of taming, by means of restrictions, the unmanageable power of the unrestricted Tarski's schema; once these restrictions are made however, deflationism is hardly very informative from a proof-theoretical point of view.

Field's latest proposal, which is more in the direction of experimenting with the fecundity of the unrestricted Tarski's schema, seems thus to require dropping the conceptual motivations of the deflationary approach as we described them. Being more liberal on the philosophical assumptions, it also provides a way to look at the complexity issue under a novel perspective. At a very general level, this aspect can even be presented as a variation of a theme of Field's (think to his remarks – end of our §3.1 – on the possibility of exercising, via the excluded middle, someone's taste to treat classically the non–classical implication *with respect to certain 'safe' domains of knowledge*). Namely, one may view the investigation on truth *really fruitful* when the latter is not regarded as an isolated notion, but it is tested on some given domain in its interplay with the other concepts which are primitively defined therein. Complexity results might then be taken as an indication of how far this 'interaction' with the primitive notions of a given domains can be pushed through.

In this direction of research, the insistence on providing *theories* we have implicitly favored by sticking to proof–theoretic facts (and thus the consequent need to take a closer look on how to circumvent the feature Field's construction shares with all the other revision processes), has first of all a methodological explanation: formal theories and proof–theoretical methods are widely accepted tools in order to deal with the domains truth–theoretical investigations are mostly experienced with (namely, mathematical domains), and approaching in a similar way the case here at stake might be convenient for the purpose of possible comparison with other pre–existing proposals.

However, additional support can be provided for that if revision processes in general become involved into our analysis, and if we regard them as possibly providing a natural generalization of theories for inductive definitions, as these have been considered so far.

Going back to the clauses we have introduced at the beginning of this section, it is known that definitions of, say, sets of integers in terms of fixed points (namely, X_γ's such that $X_\gamma = X_{\gamma+1}$), are possible in case d is *monotonic.*

Not surprisingly, this case which is the most popular one with application spanning from logic to computer science, is also the case with respect to which we have an exhaustive knowledge featuring recursion–theoretic, definability–theoretic and proof–theoretic aspects.[55] Even in the more restricted field of the formal investigation of truth, various theories inspired to seminal contribuitions on partial notions of truth preserving monotonicity, like Kripke's, have been proposed.[56]

Special cases of the non-monotonic case have been investigated as well: it is in fact known that a closure ordinal exists also in case the application of a non–monotonic operator is iterated in such a way to respect a 'cumulativity condition' (simply stating that at each stage one keeps track of what has been produced at all previous stages). There are papers both covering some recursion theory, and establishing relevant connections with the theory of admissible ordinals and sets when definitions of this sort are concerned;[57] the work by G. Jäger, partly jointly with T. Studer, instead, provides some proof theory, with a special attention to the construction of natural models for S. Feferman's systems for Explicit Mathematics via non–monotonic inductive definitions.[58]

[55] See Y. Moshovakis, *Elementary Induction on Abstract Structures*, cit. and W. Buchholz, S. Feferman, W. Pohlers and W. Sieg, *Iterated Inductive Definitions and Subsystems of Analysis: Recent Proof–Theoretical Studies*, Springer, Lecture Notes in Mathematics 450, 1981.

[56] See, e.g., A. Cantini, *Notes on Formal Theories of Truth*, «Zeit. für Math. Logik und die Grund. der Math.», 35, 1989, pp. 97–130, and, by the same author, *A Theory of Formal Truth Arithmetically Equivalent to ID_1*, «Journal of Symbolic Logic», 55, 1991, pp. 244–259, *Levels of Truth*, «Notre Dame Journal of Formal Logic», 36, 1995, pp. 185–213, A. Cantini, *Logical Frameworks for Truth and Abstraction. An Axiomatic Study*, cit., S. Feferman, *Reflecting on Incompleteness*, cit., R. Khale, *Universes Over Frege Structures*, «Annals of Pure and Applied Logic», 119, 2003, pp. 191–223.

[57] W. Richter, *Recursively Mahlo Ordinals and Inductive Definitions*, in R. O. Gandy and C. E. M. Yates, *Logic Colloquium '69*, North-Holland, 1971, pp. 273–288, P. Aczel and W. Richter, *Inductive Definitions and Reflecting Properties of Admissible Ordinals*, in J. E. Fenstad and P. G. Hinman, *Generalized Recursion Theory*, North-Holland, 1974, pp. 301–381.

[58] G. Jäger, *First-Order Theories for Non-Monotone Inductive Definitions: Recursively Inaccessible and Mahlo*, «Journal of Symbolic Logic», 66, 2001, pp.

What may an analysis specifically referring to transfinite definitions as those embodied by revision procedures add to this second direction of research? The crucial aspect of revision constructions is that, albeit non–monotone in character, they do not make use of the cumulative constraint. Each successor step is thus a singularity in the evaluation process comprised by the definition, and the usual fixed point argument do not apply. However, some relevant 'stabilization properties' of the definition are known to appear for *suitably chosen limit conditions* (the G–step in the schema of definition (3)). In particular, it is known that with respect to certain relevant choices of that sort, there exist ordinals showing that the process as a whole is *periodic*, in the sense that, from a certain point onward, blocks of stages tend to re–appear according to a period, which entails the existence of ordinals showing that the process becomes *cyclic* at some point.[59]

This specific way of presenting revision–like constructions as a natural modification of inductive definitions is the one which gives the best reason to look at these matters in the direction suggested above: to provide some theory encapsulating the defining clauses for some suitably chosen revision process, as a first step toward a general investigation of the features of the underlying definitional schema.

The results we have mentioned above then play an important role, since they may both provide important guidelines for mastering this enterprise, and also encourage to go 'through it' by keeping special attention to what the results in question make it likely to be its 'natural environment', namely the set–theoretical setting. One way in which this research may

1073–1089, and G. Jäger and T. Studer, *Extending the System T_0 of Explicit Mathematics: the Limit and Mahlo Axioms*, «Annals of Pure and Applied Logic», 114, 2002, pp. 79–101.

[59] The reader is referred to J. P. Burgess, *The Truth is Never Simple*, cit., A. Visser, *Semantics and the Liar Paradox*, in D. M. Gabbay and F. Guenthner, *Handbook of Philosophical Logic vol. IV (1st edition)*, Reidel, Dordrecht, 1989, pp. 617–706, and V. McGee, *Truth, Vagueness and Paradox*, Hackett, 1991, for a precise statement of these results, and their related proofs.

turn out to be fruitful, would be precisely the one leading to force someone envisioning some new, suitably chosen assumption on sets in order to pursue the interpretation of the theory for revision–like construction we have just suggested in a set theoretic setting. Owing to the complexity indication we have recollected so far, assumptions as such are, in terms of deductive strenght, likely to go very close to present–day limits of the metamathematical investigation.[60] Is not even unlikely,[61] that the investigation we have here suggested might contain information which would result to be crucial for those limits to be surpassed.

[60] A complete and satisfactory proof–theoretic account is available for systems which, in terms of subystems of second–order number theory, do not exceed the strenght of (Π^1_2–CA).

[61] See P. Welch, *Weak Systems of Determinacy and arithmetical Quasi–Inductive Definitions*, Unpublished notes, 2007, in particular.