Field on Revenge*

Agustín Rayo and P.D. Welch

August 16, 2006

In a series of recent papers,[1] Hartry Field has proposed a novel class of solutions to the semantic paradoxes, and argued that the new solutions are 'revenge-immune'. He has argued, in particular, that by building on a sufficiently expressive language one can get a language which is able to express its own semantic theory, including its own truth predicates and any intelligible determinacy predicates. The purpose of this note is to argue that the plausibility of Field's revenge-immunity claim depends crucially on the status of higher-order languages. We show that by availing oneself of higher-order resources one can give an explicit characterization of the key semantic notion underlying Field's proposal, and note that inconsistency would ensue if the languages under discussion were expressive enough to capture this notion.

1 Field's Proposal

For the sake of concreteness, we shall consider the version of the proposal developed in Field (2003a) and further elucidated in Field's contribution to the present volume. Start with a standard first-order language L containing the set-theoretic primitives '$\in$' and 'Set(...)', together with an arbitrary selection of non-set-theoretic predicates. On its intended interpretation, the range of L's quantifiers includes all sets, '$\in$' expresses set-theoretic membership and 'Set(...)' is true of all and only sets. In the standard, set-theoretic sense of 'model', there is no model that captures L's intended interpretation, since models are sets and, in standard set theory, no set has a transitive closure including all sets. But Field invites us to consider the next best thing to an intended model for L: the *quasi-correct* models of L. A quasi-correct model for L is a model of L in which '$\in$' expresses set-theoretic membership amongst individuals in the extension of 'Set(...)', and the extension of 'Set(...)' consists of all sets built up from the model's urelements up to some inaccessible rank (where the model's urelements are the individuals in the domain of the model but not in the extension of 'Set(...)').

So far, nothing out of the ordinary has happened. The action comes when Field extends L to a richer language L^+ containing the new one-place predicate 'Tr(...)' and the new

*Thanks to Hartry Field, Vann McGee and Graham Priest for their many helpful comments.

[1]Field (2003a), Field (2003b), Field (2004) and Field's contribution to the present volume.

two-place sentential connective '$\longrightarrow$', and uses an iterated version of the construction in Kripke (1975) to characterize what might be called the *completion* of a quasi-correct model m of L. We needn't concern ourselves with the details of Field's construction for now, except to note that if m^+ is the completion of a quasi-correct model m of L, then:

1. m^+ is a many-valued model for L^+, in which every sentence gets assigned value 1, 0 or $\frac{1}{2}$;
2. every sentence of L that is true in m gets value 1 in m^+, and every sentence of L that is false in m gets value 0 in m^+;
3. every instance of the truth-schema '$\mathrm{Tr}(\langle A\rangle) \longleftrightarrow A$' gets value 1 in m^+ (where $\ulcorner \phi \longleftrightarrow \psi \urcorner$ is an abbreviation for $\ulcorner (\phi \longrightarrow \psi) \wedge (\psi \longrightarrow \phi)\urcorner$); and
4. intersubstitution of A and $\ulcorner \mathrm{Tr}(\langle A\rangle)\urcorner$ in an arbitrary sentence of L^+ does not change the value it gets assigned by m^+.

Moreover, when validity in L^+ is defined as preservation of value 1 under all uniform substitutions of non-logical vocabulary in the completion of a quasi-correct model of L, one gets a weakening of classical logic that Field calls LCC (short for 'the Logic of Circularly Defined Concepts'). LCC fails to validate excluded middle or the equivalence between $\ulcorner \phi \longrightarrow \psi \urcorner$ and $\ulcorner \neg\phi \vee \psi \urcorner$, but is strong enough to be interesting and has a number of attractive features. And an immediate consequence of clause 2 above is that LCC behaves classically for the special case of formulas not containing 'Tr(...)' or '$\longrightarrow$'.

These results put Field in a position to explain how one might acquire a concept of truth for L^+ that would allow one to accept every instance of the truth-schema. One starts out speaking L and reasoning classically, and sets out to speak L^+. The first step is to acquire a novel understanding of the logical vocabulary:

> One makes the assumption that the logical vocabulary in L^+ is to be understood on the basis of its role in LCC-inference, and uses the concept of LCC-validity to develop such an understanding.

Since LCC-inference behaves classically when restricted to formulas of L there is no risk that one's novel understanding of the standard logical vocabulary will force one to give up inferences one previously endorsed. And since the notion of LCC-validity is characterized within standard-set theory, which is expressible in L, there is no risk of circularity.

The next step is to acquire an understanding of the new predicate 'Tr(...)'. This can be done as follows:

> One resolves to accept every L^+-instance of '$\mathrm{Tr}(\langle A\rangle) \longleftrightarrow A$' , and makes the assumption that 'Tr(...)' is to be understood on the basis of its role in this schema.

But how does one know that this won't lead to inconsistency? It is here that Field's *pièce de résistance* comes in: one can use the observation that any quasi-correct model of L has a completion to show that the result of enriching ZFC with every instance of the truth-schema has no untoward LCC-consequences. And since, given an arbitrary quasi-correct model, Field's construction can be carried out entirely within standard set-theory, it can be carried out using the expressive resources of L and therefore the expressive resources of a classically-behaving fragment of L^+. So one can convince oneself that ZFC plus every instance of the truth-schema has no untoward LCC-consequences without ascending to a strictly richer metalanguage. (On the other hand, one can only prove that L has a quasi-correct model by assuming a theory at least as strong as ZFC plus a large-cardinal hypothesis. So one's warrant for the claim that one's preferred system of set-theory plus every instance of the truth-schema is LCC-consistent presupposes a warrant for the claim that ZFC plus a suitable large-cardinal hypothesis is classically consistent.)

2 A small interlude

Field's *pièce de résistance* shows that ZFC plus every instance of the truth-schema has no untoward LCC-consequences. But, as Vann McGee pointed out to one of us in conversation, this does not immediately guarantee that one won't be able to derive a contradiction in LCC from true sentences of L and instances of the truth-schema. For it is consistent with ZFC that no quasi-correct model of L assigns every sentence in L its intended truth-value. And if there is no such model, no completion of a quasi-correct model is such that every true sentence of L and every L^+-instance of the truth-schema is assigned value 1.

Here is an analogy. Suppose there are precisely 13 inaccessibles. Then if I_{13} is a sentence of L stating that there are 13 inaccessibles, I_{13} is false according to any quasi-correct model of L, but true *simpliciter*. So although the existence of quasi-correct models of L guarantees that ZFC plus the negation of I_{13} is consistent, it does not guarantee that one won't be able able to derive a contradiction from true sentences of L and the negation of I_{13}.

Field's *pièce de résistance* makes it seem extremely plausible that one won't be able to derive a contradiction in LCC from true sentences of L and instances of the truth-schema. But it is important to be clear that—in the absence of a principle guaranteeing the existence of a quasi-correct model of L assigning every sentence in L its intended truth-value—all Field has given us is a plausibility argument. An immediate consequence of the result in section 5 is that one can transform Field's plausibility argument into a proof by availing oneself of higher-order resources.

3 Revenge

Field has a plausible argument for the claim that the truth-schema won't lead to inconsistency in LCC. We would like to suggest, however, that he only gets consistency because the language under discussion is unable to express a key semantic notion. We will see that the notion of an intended interpretation for L^+ can be characterized using higher order resources, and that inconsistency would immediately ensue if the intended interpretation of L^+ was expressible in L^+.

Consider a warm-up case. Suppose we took our initial object language to be the language of first-order arithmetic, L_A, rather than L. Whereas there is no model corresponding to the intended interpretation of L, it is easy to construct a model, m_A, that captures the intended interpretation of L_A: m_a is just the standard model of arithmetic, $\mathbb{N}$. Field's construction can be applied to m_A just as it was applied to quasi-correct models of L. What one gets is a many-valued model m_A^+ for L_A^+ (which is the result of extending L_A with'$\mathrm{Tr}(\ldots)$' and '$\longrightarrow$'). But since m_A is the intended model of L_A, one can expect m_A^+ to count as the intended model of L_A^+. In other words, one can expect the value m_A^+ assigns a sentence of L_A^+ to be its 'real' truth-status. One might say, for example, that a sentence of L_A^+ ought to be *accepted* just in case it is assigned the value 1 by m_A^+, and *rejected* just in case it is assigned value 0 or $\frac{1}{2}$. It is a consequence of Field's construction that the value of a negation is always 1 minus the value of its negatum. So whereas sentences of L_A^+ whose m_A^+-value is $\frac{1}{2}$ are such that both they and their negations ought to be rejected, sentences of L_A^+ whose m_A^+-value is 0 are such that their negations ought to be accepted even though they themselves ought to be rejected.

This yields all the right results. Every arithmetical truth, every instance of the truth-schema for L_A^+ and all of their LCC-consequences are assigned m_A^+-value 1, so they ought to be accepted and their negations ought to be rejected. The Liar sentence Q is assigned m_A^+-value $\frac{1}{2}$, as are $\ulcorner\neg Q\urcorner$, $\ulcorner\mathrm{Tr}(\langle Q\rangle)\urcorner$, $\ulcorner\neg\,\mathrm{Tr}(\langle Q\rangle)\urcorner$, and the corresponding instances of excluded middle: $\ulcorner Q \vee \neg Q\urcorner$ and $\ulcorner\mathrm{Tr}(\langle Q\rangle) \vee \neg\,\mathrm{Tr}(\langle Q\rangle)\urcorner$. So one can say—as Field does in discussing the set-theoretic case—that "one must reject the claim that $\mathrm{Tr}(\langle Q\rangle)$ and also reject the claim that $\neg\mathrm{Tr}(\langle Q\rangle)$ [...] We must likewise reject the corresponding instance of excluded middle". (Chapter X, p. Y). And, of course, the *pièce de résistance*: one can use m_A^+ to show that sentences that ought to be accepted will never have untoward LCC-consequences. (One can't carry out the proof in L_A, since m_A^+ can't be characterized in L_A: it is a consequence of Tarski's Theorem that there is no formula $\phi(x)$ of L_A such that a sentence S of L_A^+ has m_A^+-value 1 just in case $\phi\langle S\rangle$ is true.)

An account of truth for the language of arithmetic based on this idea would have many virtues. But it should be clear that revenge-immunity is not one of them. Say that a language is *revenge-immune* just in case it is able to express its own semantic theory, including its own truth predicate and any intelligible determinacy predicates. Then it should be clear that L_A^+ is not revenge-immune. For inconsistency would immediately ensue if L_A^+ was able to express the notion of acceptability—or, equivalently, the notion of having m_A^+-value 1.[2] So although it is true that the result of adding every L_A^+-instance of

[2] *Proof:* Working within a classical metalanguage, assume there is an open formula '$One(x)$' of L_A^+ such that $\ulcorner One(\langle A\rangle)\urcorner$ gets m_A^+-value 1 just in case A gets m_A^+-value 1 and use Gödelian techniques to find a sentence $Q^\star$ of L_A^+ such that $\ulcorner Q^\star \longleftrightarrow \neg One(\langle Q^\star\rangle)\urcorner$ gets m_A^+-value 1. The fact that $\ulcorner Q^\star \longleftrightarrow \neg One(\langle Q^\star\rangle)\urcorner$ gets m_A^+-value 1 guarantees that $Q^\star$ gets m_A^+-value 1 just in case $\ulcorner\neg One(\langle Q^\star\rangle)\urcorner$ gets m_A^+-value 1. But the way in which '$One(x)$' was introduced guarantees that $\ulcorner\neg One(\langle Q^\star\rangle)\urcorner$ gets m_A^+-value 1 just in case $Q^\star$ fails

the truth-schema to the standard axioms of arithmetic is LCC-consistent, this is merely because L_A^+ is unable to express the notion of having m_A^+-value 1.

So why couldn't one generate an analogous revenge problem for the set-theoretic case? The crucial observation is that the argument in the arithmetical case makes use of the fact that m_A is an intended model for the language of arithmetic, and, as emphasized above, the language of set-theory has no intended models. (Had m_A been chosen to be an unintended model for L_A, there would have been no reason for thinking that its completion should count as an intended model for L_A^+, and the proof in footnote 2 would have been open to a charge of equivocation by failing to distinguish between being assigned value 1 by m_A^+ and having 'real' value 1.)

What would happen if one tried to apply a version the revenge argument to the set-theoretic case on the basis of a quasi-correct model for L, rather than an intended model? Assume one has defined in L a quasi-correct model m of L, and let m^+ be the completion of m. It is a consequence of Tarski's Theorem that no model of L definable in L assigns every sentence of L its intended truth-value. So thinking that m^+ generates a revenge problem for L^+ would be like thinking that one violates Tarski's Theorem by characterizing the notion of truth-according-to-m within the language of set-theory. Tarski's Theorem is not violated because truth-according-to-m is not genuine truth; similarly—as Field is careful to point out—revenge problems are averted because having m^+-value 1 does not correspond to genuine acceptability.

Assume one has defined a quasi-correct model m of L within the language of first-order set theory, and let m^+ be the completion of m. As Field is careful to point out, the notion of having m^+-value 1 can be defined within L, since L includes the language of first-order set-theory. So the notion of having m^+-value 1 is expressible in L^+. But thinking that m^+ generates a revenge problem for L^+ would be like thinking that one violates Tarski's Theorem by characterizing the notion of truth-according-to-m within the language of set-theory. Tarski's Theorem is not violated because truth-according-to-m is not genuine truth; similarly, revenge problems are averted because having m^+-value 1 does not correspond to genuine acceptability.

Nevertheless, there might be quasi-correct models of L (not definable in L) that assign every sentence of L its intended truth-value. (This can occur, for example, if there are κ inaccessibles for which $V_\kappa \prec V$.) Suppose there are indeed such models, and let κ_0 be least so that $m = V_{\kappa_0}$ is such. Then one might expect m^+ to assign every sentence of L^+ its 'real' truth-status, and one might argue against Field's revenge-immunity claim by suggesting that although it is true that the result of adding every L^+-instance of the truth-schema to the standard axioms of arithmetic is LCC-consistent, this is simply because L^+ is unable to characterize the notion of m^+-value 1. (The fact that m cannot be defined ZFC in the language L alone is presumably no impediment to the intelligibility of the notion of m^+-value 1, since m—and therefore m^+—is definable in the language that results from enriching L with a Tarskian truth-predicate for L.)

This argument, such as it is, against Field's revenge-immunity claim depends both on

---

to get m_A^+-value 1.

the assumption that there is a quasi-correct model m of L that assigns every sentence of L its intended truth-value and on the assumption that the completion of m ascribes every sentence of L^+ its 'real' truth-status even though m is not the intended model of L. In the following section we will see that no such assumptions are necessary when higher-order resources are brought on board.

4 SO-Models

By availing oneself of higher-order resources, it is possible to characterize the intended interpretation of L. Details are supplied in Rayo and Uzquiano (1999), but the basic idea is straightforward. In standard model-theory one says of an individual—a set with certain properties—that it is a model. But by availing oneself of higher-order resources one could also say of some individuals—some ordered-pairs with certain properties—that they (jointly) form an SO-model (short for 'second-order model'). If the Gs form an SO-model, one says that x is in the 'domain' of the Gs just in case $\langle `\forall', x\rangle$ is one of the Gs, and one says that the atomic predicate-letter P is true of x according to the Gs just in case $\langle P, x\rangle$ is one of the Gs. Truth and satisfaction can then be characterized along familiar lines. For any standard model m there are some things that form an SO-model with the same domain as m and the same interpretations for atomic predicates as m (and therefore the same truths as m). But the big advantage of SO-models is that they allow for domains too big to form a set. There are, in particular, some individuals that together form the intended SO-model for L. Let us call them 'the M's'. (Something is one of the M's just in case it is either an ordered-pair of the form $\langle `\forall', x\rangle$ (for $x$ an arbitrary object), an ordered-pair of the form $\langle `\text{Set}', \alpha\rangle$ (for α a set), an ordered-pair of the form $\langle `\in', \langle\alpha, \beta\rangle\rangle$ (for α and β sets such that $\alpha \in \beta$), or an ordered-pair of the form $\langle \ulcorner P_i^n \urcorner, \langle x_1, \ldots, x_n\rangle\rangle$ (for $\ulcorner P_i^n \urcorner$ true of $x_1, \ldots, x_n$)).

In order to generate revenge problems for Field, it is not enough to show that the M's capture the intended interpretation of L, one must also find a way of extending the M's to an intended interpretation of L^+ by applying an analogue of Field's construction. But, as we shall see in section 5, Field's construction can be emulated in a higher-order setting. Let the M$^+$'s be the result of applying the higher-order version of Field's construction to the M's. The M$^+$'s form a many-valued SO-model for L^+. And since the M's are the intended model of L, one can expect the M$^+$'s to count as the intended model for L^+. In particular, one can expect the value that the M$^+$'s assign a sentence of L^+ to be its 'real' truth-status. As in the arithmetical case, one might say that a sentence of L^+ ought to be *accepted* just in case it is assigned the value 1 by the M$^+$'s, and *rejected* just in case it is assigned value 0 or $\frac{1}{2}$. It will still be a consequence of the construction that the value of a negation is always 1 minus the value of its negatum. So whereas sentences of L^+ that get assigned the value $\frac{1}{2}$ by the M$^+$'s are such that both they and their negations ought to be rejected, sentences of L^+ that get assigned the value 0 by the M$^+$'s are such that their negations ought to be accepted even though they themselves ought to be rejected.

As in the arithmetical case, this yields all the right results. Every truth of L, every instance of the truth-schema for L^+ and all of their LCC-consequences are assigned value

1 by the M^+'s, and therefore ought to be accepted and have negations that ought to be rejected. The Liar sentence Q is assigned value $\frac{1}{2}$ by the M^+'s, as are $\ulcorner\neg Q\urcorner$, $\ulcorner\mathrm{Tr}(\langle Q\rangle)\urcorner$, $\ulcorner\neg\,\mathrm{Tr}(\langle Q\rangle)\urcorner$, and the corresponding instances of excluded middle: $\ulcorner Q \vee \neg Q\urcorner$ and $\ulcorner\mathrm{Tr}(\langle Q\rangle) \vee \neg\,\mathrm{Tr}(\langle Q\rangle)\urcorner$. So one can say, with Field, that "one must reject the claim that $\mathrm{Tr}(\langle Q\rangle)$ and also reject the claim that $\neg(\mathrm{Tr}\,\langle Q\rangle)$ [...] We must likewise reject the corresponding instance of excluded middle". Finally, one can use the M^+'s to show that sentences that ought to be accepted will never have untoward LCC-consequences.

As in the arithmetical case, the resulting picture has many virtues. But revenge-immunity is not one of them. Inconsistency would immediately ensue if L^+ was able to express the notion of acceptability—or, equivalently, the notion of being assigned value 1 by the M^+'s. (The proof is exactly analogous to that in footnote 2.) So although it is true that the result of adding every L^+-instance of the truth-schema to the standard axioms of set theory is LCC-consistent, this is merely because of L^+'s is unable to express the notion of being assigned value 1 by the M^+'s.[3]

Needless to say, the argument against revenge-immunity in the set-theoretic setting will be a non-starter unless one is willing to countenance the legitimacy of higher-order languages. The status of higher-order languages is the subject of intense debate. We cannot hope to address this debate here.[4] All we wish to show is that the plausibility of Field's revenge-immunity claim is sensitive to the outcome.

5 The Construction

In order to generate revenge problems for Field, it is not enough to show that the M's capture the intended interpretation of L, one must also find a way of extending the M's to an intended interpretation of L^+ by applying an analogue of Field's construction. In order to do this we sketch the development of a fragment of second order set theory over V, the class of all sets. (Talk of classes is a notational convenience: first-order quantification over classes is to be thought of as a syntactic abbreviation for second-order quantification over individuals; similarly, first-order quantification over collections of classes is to be thought of as a syntactic abbreviation for third-order quantification over individuals.)

At the heart of Field's construction over a suitable first-order model m is a classical recursive process along an initial segment of the ordinals. The clauses of this recursion interleave a hybrid constuction of obtaining a Kripkean fixed point using the strong Kleene scheme of partial logic with an evaluative process for $\longrightarrow$ which encapsulates a history of what has happened at previous stages. At each ordinal stage, (i) all sentences involving Tr are reset to have value $\frac{1}{2}$; (ii) the semantic value of sentences involving $\longrightarrow$ are calculated according to the aforementioned process; (iii) finally a new fixed point *à la* Kripke is

[3]For further discussion of this point, see Priest's contribution to this volume.

[4]For the conservative side of the debate see Quine (1986) ch. 5, Resnik (1988), Parsons (1990) and Linnebo (2003), among others. For the liberal side of the debate see Boolos (1984), Boolos (1985a), Boolos (1985b), McGee (1997), Hossack (2000), McGee (2000), Oliver and Smiley (2001), Rayo and Yablo (2001), Rayo (2002), Williamson (2003) and Rayo (forthcoming), among others.

computed. The latter process assigns values to sentences involving Tr (but does not alter the values of the binary operator $\longrightarrow$; these remain fixed through the process of calculating the next fixed point). At the beginning of each ordinal stage α say, during (ii), essentially a Σ_2-recursive clause is invoked: the semantic value of $A \longrightarrow B$: the value here depends on whether *there exists* a previous stage β so that *for all* subsequent stages γ before α something happens about the valuations of A and B at those stages γ.

It is this latter clause that gives the construction its essential flavour, and its overall complexity. Nevertheless, as m is a set, there will be some least *acceptable ordinal* $\Delta_0(m)$ at which the whole process stabilises out and starts to cycle. Field (2003a) demonstrates the existence of such acceptable ordinals, and in Welch (2003) several equivalent characterisations of such are given. Essentially Welch (2003) shows that these characterisations can be obtained by a demonstration that over $m = \mathbb{N}$, the standard model of arithmetic, the set of sentences that have "ultimate" semantic value 1 (*i.e.* will receive semantic value 1 at some point never to be later changed) is recursively isomorphic to Herzberger's set of *"stable truths"* obtained from a single revision sequence starting from a distribution of semantic value 0 to all sentences. Other characterisations of this set were already known, and yielded a computation of Field's least accetptable ordinal $\Delta_0(\mathbb{N})$.

That article also mentioned that the whole construction over $\mathbb{N}$ could be performed in second order number theory (the "Second Demonstration"); although it urged the reader not to do so (since the possibility of doing so was given explicity by the argument above - the "Third Demonstration" and the details would be wearisome) it is in fact this "second demonstration" that we must perform here, albeit translated to the second order set-theoretical arena, rather than the number-theoretic one.

As V contains all ordinals we must elaborate a theory of *wellorderings* given by class terms that are of sufficient length to allow us to prove that Field's recursive construction can be emulated along these *wellorderings*. In case the reader is a little queasy about the notion of such orderings, we could point out that it is easy in standard first order ZF set theory to define orderings that have the apparent "order type" '$\mathrm{On} + \mathrm{On}$' where On denotes the class of all ordinals: one simply defines a class of pairs $\langle \alpha, i \rangle$ for $\alpha \in \mathrm{On}, i \in \{0, 1\}$ and defines $\langle \alpha, i \rangle <' \langle \beta, j \rangle \longleftrightarrow i < j \vee (i = j \wedge \alpha < \beta))$ (where $<$ has its usual meaning). It is simply that such order-types are not *set-like*, that is they do not have initial segments that are sets. Nevertheless with care, set theorist can, and do, use such orderings. Similar definitions are immediate for $\mathrm{On} \times \mathrm{On}$, $\mathrm{On}^{\mathrm{On}}$ etc., etc. However the ordering type required in Field's construction cannot be given by any first order class terms definable over V. Thus, for example, the construction cannot be done in vNGB set theory. Indeed we shall see that instances of Π^1_3 Comprehension in second order set theory are needed (which we shall call Π^1_3-CA by analogy with that of subsytems of analysis (see Simpson (1999))).

In the following L will be the standard countable first order language for set theory with equality and set membership symbol ϵ as the sole binary predicates, and without constants. We shall add constants to L, one for each set $x \in V$. Developing the syntax for this language using sets as codes, can be done in a weak set theory (and so in ZFC) (see, for example Devlin (1984) Ch.I.9 for a standard account). We shall call this language L_V; done over over $(V, \in)$, this yields a class term for the elements of this language: thus

$L_V \subseteq V$.

These languages augmented with the predicate symbol Tr and a binary conditional relation $\longrightarrow$ will be called L^+, and L_V^+ respectively (the latter language is also then given by a simply defined class term over $(V, \in)$).

We sketch the adaptation of the first stage of Field's construction (now over $(V, \in)$). This first starts out with assigning semantic values of $\frac{1}{2}$ to all atomic sentences of L_V^+ of the form $\mathrm{Tr}(u)$ where u is (a set coding) a sentence of L_V. Similarly all sentences of the form $A \longrightarrow B$ where $A, B \in L_V$ are also set to value $\frac{1}{2}$. The rest of the clauses for this first stage are those needed to construct in a familiar manner the first strong Kleene fixed over the structure $(V, \in)$in the language L_V^+. Those sentences assigned a semantic value of 1 then form a proper subclass of L_V^+. This class is not given by a term *definable* over $(V, \in)$ but is the result of an *inductive* second order process over $(V, \in)$. (Similarly for the classes of sentences with semantic values $0, \frac{1}{2}$.) The use of the language L_V^+ enriched with constants for all sets allows a simple substitutional quantification clause for the working out the values of $|\forall v A(v)|_{0,\sigma}$ and $|\exists v A(v)|_{0,\sigma}$ (to use Field's notation for the semantic values calculated at the σ'th substage of this - the first full stage of the cumulative process of computing Strong Kleene fixed points.)

Supposing that $(X^1, X^0)$is a pair of classes contained in L_V^+ with X^1 those receiving value 1, X^0 receiving value 0 (and the rest by default $\frac{1}{2}$), at some substage. We can view one step of this process as a (Strong Kleenian) *jump operation*: $j((X^1, X^0)) = (Y^1, Y^0)$ which delivers the extensions of those classes receiving the respective semantic values at the next stage: $u \in Y^1$ (and so in the extension of Tr) at the next substage, if at the previous substage certain "truth conditional" clauses are fulfilled. (We'll set $j^1((X^1, X^{-0})) = Y^1$ and $j^0((X^1, X^0)) = Y^0$.) For any particular u these are elementary conditions on $(V, \in, (X^1, X^0))$. Overall we may say that the relations

$$\text{“}u \in j^1((X^1, X^0))\text{”; “}u \in j^0((X^1, X^0))\text{”}$$

whilst not elementary for disjoint classes $X^1 \cap X^0 = \varnothing$, are Δ_1^1 over $(V, \in)$. (A Π_1^1 relation $\mathcal{R}(v, Z)$ over $(V, \in)$ is one of the form $\mathcal{R}(v, Z) \iff \forall U \varphi(v, Z, U)$ where φ is elementary in the language L_V with additional second order variables $\dot{Z}, \dot{U}$. A Σ_1^1 relation is the complement of a Π_1^1 relation, and a relation is Δ_1^1 if both it and its complement can be written in Π_1^1 form. For the case in hand the relations are not quite elementary since u could be the code of a sentence of arbitrary complexity in the usual Levy hierarchy of classification of formulae, but the relations are not far from elementary.) In Field's notation, if (X^1, X^0) are the classes of sentences assigned semantic value 0/1 at stage σ then $|\mathrm{Tr}(u)|_{\sigma+1} = 1 \iff u \in j^1((X^1, X^0))$, etc.

In short this first stage of establishing the Strong Kleene fixed point in L_V^+ is a *monotone inductive* process *over* $(V, \in)$. In particular we can think of this couched in terms of the extension of the theory of inductive definability to a theory of inductive *second order* relations as adumbrated in Moschovakis (1974) Ch. 6.

Similarly the second order relation $\mathcal{U}(X, Y)$:

$$\text{“}X \cap Y = \varnothing \wedge j^1(X \cap Y) \subseteq X \wedge j^0(X \cap Y) \subseteq Y\text{”}$$

is also Δ^1_1.

Consequently the first fixed point (A^1_0, A^0_0) where (in Field's notation $u \in A^1_0 \Longleftrightarrow |u|_0 = 1$ and further $u \in A^1_0 \Longleftrightarrow \mathrm{Tr}(u) \in A^1_0$, and so on) is Π^1_1-definable:

$$u \in A^+ \Longleftrightarrow \forall X^1 \forall X^0 (\mathcal{U}(X^1, X^0) \Longrightarrow u \in X^1)\text{”}.$$

We thus have that both A^1_0 and A^0_0 are both Π^1_1 and also inductive over $(V, \in)$. (There should be a word of warning here: over $\mathbb{N}$ Π^1_1 and (positive) inductive relations coincide, but $\mathbb{N}$ is a special case, and over other structures positive elementary inductive relations are Π^1_1, but the converse may fail in general, so not all results will generalise.) Of course the "induction" that is implicit in the above would in the set-sized case have a particular length (again in Field's terminology, that $|u|_0$ appearing in the above is strictly $|u|_{0,\Omega}$ for some "sufficiently large" ordinal Ω).

As already adverted, in our context a wellordering of sufficient length along which to run the Kripkean construction is beyond any length of a first order definable over $(V, \in)$wellordering $<$ of On. We further need also to define the semantic values of formulae of the form $A \longrightarrow B$, *after* each fixed point has been reached. In particular this whole process of finding fixed points and evaluating "conditionals" will have to be repeated for more stages than there are ordinals, for a particular liminf limit rule to be applied at "limit" stages. We thus instead resort to a weak second order set theory.

We reserve upper case letters T, W, U,to informally denote classes (as we have been doing), with X_i,etc to be class variables. Sets t, w, u will be lower case, with x_i etc being set variables. An $\mathcal{L}^2$ *structure* is as follows:

$\mathfrak{M} = (V, S_M, \in)$

where $S_M \subseteq \mathcal{P}(V)$ is a collection of *classes*, and is used to interpret the second order variables of $\mathcal{L}^2$. If $\mathcal{B}$ is any subclass of $V \cup S_M$ then $\mathcal{L}^2_{\mathcal{B}}$ is the language of $\mathcal{L}^2$ augmented by constants from $\mathcal{B}$. A class $W \subseteq V$ is *definable over* $\mathfrak{M}$ *using parameters from* $\mathcal{B}$ if there exists a formula $\varphi(v_0) \in \mathcal{L}^2_{\mathcal{B}}$ so that $W = \{w \in V | M \models \varphi(w)\}$. We take as axioms:

Definition 1 Γ-Comprehension Axioms*(Γ-CA) comprise:*

(i) *The usual* ZFC *axioms for sets;*
(ii) *A second order induction axiom:*

$$\forall \alpha (\forall \beta < \alpha (\beta \in X \longrightarrow \alpha \in X)) \longrightarrow \forall \alpha (\alpha \in X))$$

(iii) Γ-*Comprehension scheme (where* $\Gamma \subseteq \mathcal{L}^2$ *is a class of second order formulae)*

$$\exists X \forall y (y \in X \longleftrightarrow \varphi(y))$$

for any $\varphi \in \Gamma$.

In the above Γ will usually restricted to be of the form Σ^1_1, Π^1_n etc., etc. We shall interpret the set variables as always ranging over the *standard universe* V: there will thus be no non-standard models: continuing the analogy of second order number theoretic, we are thus assuming all models are 'ω-models'. Continuing this analogy:

Definition 2 $\mathfrak{M} = (V, S_M, \in)$ *is a* Σ^1_k-correct model (of set theory) *iff whenever* φ *is a* Σ^1_k*-sentence of* $\mathcal{L}^2$*, then* φ *is true if and only if* $\mathfrak{M} \models \varphi$.

Of course the 'φ is true' here has to be interpreted relative to some ambient domain of discussion containing $\mathfrak{M}$. This domain will later be taken to be a larger model $\widetilde{\mathfrak{M}} = (V, S_{\tilde{M}}, \in)$ with $S_{\widetilde{M}} \supseteq S_M$. Without specifying it further at the moment, we consider the following discussion to be developed within this model.

We note that the assertion that some $W \in S_M$ for which $W \subset (\mathrm{On} \times \mathrm{On})$ is a class of ordered pairs which is a *linear ordering* is first order (or "elementarily") definable in $\mathfrak{M}$ (it is in '$\Pi^1_0(\mathfrak{M})$'). (We merely have to assert that the class of pairs satisfies the usual requirements of transitivity, anti-symmetry, and trichotomy of a (strict) total order, which are simply expressed using universal quantifiers over On.) An assertion concerning "well-ordering" however is still just an elementary quantification. (For any $W \in S_M$ $\mathfrak{M} \models$"W is a wellordering of On" can be expressed as "$\forall \tau \in \mathrm{On}\, W \cap \tau \times \tau$ is wellordered".) Our models are therefore automatically correct about which classes wellorder (a subclass of) the ordinals. Thus for such models the notion of being a class wellordering is absolute. We further claim that that the definition of, for example, Σ^0_1-satisfaction over second order models $\mathfrak{N}$ can itself be given in a Σ^0_1 fashion, by basically the same reasoning as for the first order case.

The narrative now is developed parallel to that of second order number theory, assuming as a base theory our structures $\mathfrak{M}$ model first order, or "elementary" comprehension ("ECA").

Note that a necessary and sufficient condition to be a Σ^1_1-model is that for any $X \in S_M$ that the complete $\Sigma^1_1(X)$ definable class over $\mathfrak{M}$ also be in S_M. Moreover:

Lemma 1 *Suppose* $\mathfrak{M} = (V, S_M, \in)$ *is a model of ECA. Then the following are equivalent:*
(i) $\mathfrak{M}$ *is a* Σ^1_1*-correct model of* Π^1_1-CA *;*
(ii) For any $X \in S_M$*, the complete* $\Sigma^1_1(X)$ *class is in* S_M.

Taking $\mathfrak{N} = (V, \varnothing, \in)$, we may look at the least collection of classes containing those first order definable over $\mathfrak{N}$, and then closing under the operation of taking for any class X, the complete $\Sigma^1_1(X)$ class. Let S be the resulting collection of classes. We then have:

Lemma 2 $\mathfrak{M} = (V, S, \in)$ *is the minimum* Σ^1_1*-model of* Π^1_1-CA. *Similarly for any class* X *there is a minimum* Σ^1_1*-correct model of* Π^1_1-CA*,* $\mathfrak{M}_X = (V, S_{M_X}, \in)$ *with* $X \in S_{M_X}$.

Definition 3 *An* ordinal coded Σ^1_1-correct model *is a class* $W \subset \mathrm{On} \times V (\subset V)$ *which codes a* Σ^1_1*-correct model* $\mathfrak{M} = (V, S_M, \in)$ *via* $S_M = \{(W)_\alpha | \alpha \in \mathrm{On}\}$ *where* $(W)_\alpha = \{x | (\alpha, x) \in W\}$

We can then think of certain Σ^1_1-correct models $\mathfrak{M}$ as themselves just classes, if S_M is so enumerable by some class $W = W(\mathfrak{M})$.

Lemma 3 Π^1_1-CA $\Longleftrightarrow$*For all X there is an ordinal coded Σ^1_1-correct model $\mathfrak{M}_X$ with $X \in S_M$.*

We shall need certain second order schemes of recursion available; let us call these *class recursions*: the idea is that we suppose we have second order formula $\varphi(x, X) \in \mathcal{L}^2$ (with possibly other set or class parameters), then given a wellordering $W \subset V \times V$, for each $x \in \mathrm{Field}(W)$ we shall associate a class Y_x by a recursion along W: if Y_z has already been defined for $z <_W x$ then $Y^x =_{\mathrm{df}} \{(u, v)|u \in Y_v \wedge v <_W x\}$, and then $Y_x =_{\mathrm{df}} \{q|\varphi(q, Y^x)\}$.

Definition 4 *The scheme of* Π^1_k-class recursion *(Π^1_k-REC) comprises all instances of the following where $\varphi(x, W) \in \Pi^1_k$ (possibly with other set and class parameters):*

$$\forall W(\mathrm{WO}(W) \longrightarrow \exists Y H_\varphi(W, Y))$$

where H_φ says that Y is the class of all pairs (u, x) such that $x \in \mathrm{Field}(W)$ and $\varphi(u, Y^x)$ where $Y^x =_{\mathrm{df}} \{(u, v)|(u, v) \in Y \wedge v <_W x\}$.

As Welch (2003) argues in the Third Demonstration, that what one needs for Field's construction over $\mathbb{N}$ to succeed is a "Σ_2-extendible ordinal": an ordinal large enough that it enjoys a certain amount of reflection in Gödel's constructible hierarchy L built over $\mathbb{N}$. Proof-theoretically, in this Second Demonstration we need a sufficiently strong theory so that we can prove the existence of certain ordinal coded Σ^1_1-correct models with similar sufficiently strong reflection properties: it is sufficient to have that there is at least one such Σ^1_1-correct model $\mathfrak{M}$, which enjoys Σ^1_3-reflection into an ordinal coded submodel $\mathfrak{N}$ (which perforce will be Σ^1_1-correct), which we shall write as: $\mathfrak{N} \prec_{\Sigma^1_3} \mathfrak{M}$. More formally:

Definition 5 *(i) $\mathfrak{N}$ is a* submodel *of $\mathfrak{M}$ if $S_N \subseteq S_M$;*

(ii) $\mathfrak{N}$ is a Γ-submodel *of $\mathfrak{M}$ if for all sentences $\varphi \in \Gamma$ with parameters from $\mathfrak{N}$ we have $\mathfrak{N} \models \varphi \Longleftrightarrow \mathfrak{M} \models \varphi$.*

By analogy with Lemma 3 above we have at the third level:

Lemma 4 Π^1_3-CA *proves that for every X there is an ordinal coded Σ^1_3-correct model $\mathfrak{M}_X$ with $X \in S_M$*

We now apply Lemma 4 twice: the first time to get an ordinal coded Σ^1_3-correct model $\mathfrak{N}$ with $X \in S_N$, and then a second time to get a model $\mathfrak{M}$ with a code for $\mathfrak{N}$ in S_M. As both models are Σ^1_3-correct we obtain the right amount of reflection:

Lemma 5 Π^1_3-CA *proves that for every X there are ordinal coded models $\mathfrak{N}, \mathfrak{M}$ with $\mathfrak{N} \prec_{\Sigma^1_3} \mathfrak{M}$ and $X \in S_N$.*

If we then assume our ambient universe $\tilde{\mathfrak{M}} = (V, S_{\tilde{M}}, \in)$ is a model of Π^1_3-CA, then the conclusion of Lemma 5 holds. Take $X = V$ and we shall have a pair of models $(\mathfrak{N}_0, \mathfrak{M}_0)$ as in Lemma 5. Let us fix our attention on this pair.

Given a distribution of semantic values (X^1, X^0) for an L^+ model $\mathfrak{N}$, by Π^1_1-CA we can find the strong Kleene fixed point class for this distribution. To calculate the ultimate acceptable semantic values we need to iterate this process along a wellordering given to us by Π^1_3-CA, namely the wellorderings of the second order model $\mathfrak{M}$. The successor stages are given by calculating the next Kripkean fixed point - and so are given by a Π^1_1-recursion. However at limit points of the wellordering $W = (\text{Field}(W), <_W)$, we see that if say $u \in \text{Field}(W)$ is a $<_W$-limit then evaluations of $|A \longrightarrow B|_u$ are given by the following

$$|A \longrightarrow B|_u = \begin{cases} 1 & \text{iff } (\exists w <_W u)(\forall v \in [w,u]_W)(|A|_v \leq |B|_v), \\ 0 & \text{iff } (\exists w <_W u)(\forall v \in [w,u]_W)(|A|_v > |B|_v) \\ \frac{1}{2} & \text{otherwise} \end{cases}$$

(Actually the above clauses define $|A \longrightarrow B|_u$ for *any* $u \in \text{Field}(W)$ but for u the successor of u_0 the evaluation is just elementary in the Kripkean fixed point evaluations at stage u and u_0, and hence could be considered as part of an overall Π^1_1 recursion, were it not for the fact that it is at the *limit* stages where the above definition has all its bite.) As can be seen by the further outer pair of alternating $\exists\backslash\forall$quantifiers in this definition by cases, we shall have here a Σ^1_3-recursive clause.

We then finally have:

Lemma 6 Π^1_4-CA *proves that the Fieldian construction starting with V - the "intended model" of set theory - succeeds,* i.e. *that there is a model V^+.*

Proof: Assume our ambient universe is a model of Π^1_4-CA: this is sufficient to iterate Π^1_3-CA along any wellorder; this means that we have Π^1_3-REC. In our case we can take any wellorder in S_{M_0} that is longer than the supremum of those in our chosen ordinal coded Σ^1_3-correct $\mathfrak{N}_0$ from Lemma 5. We thus have, in particular, if $W \in S_{M_0}$ is a wellorder of rank greater than anything in S_{N_0}, and if $u \in \text{Field}(W)$ has rank the supremum of ranks of wellorderings in S_{N_0}, that the values of the u'th iterate along W is that of an *acceptable point* in the Fieldian construction (and that of the rank of $\mathfrak{M}$ would be the second acceptable point). Setting $\|A\| = |A|_u$ then, we get the ultimate semantic values for each $A \in L^+$,and have thus constructed the intended V^+. Q.E.D.

Remark 1 Π^1_4-CA is not best possible in the last lemma. If we had assumed only Π^1_3-CA, and a kind of formal development of a constructible hierarchy over V as detailed for second order arithmetic in Simpson (1999) VII.4, then using absolutness and relativisation arguments, we could find a model of $\Pi^{1-\text{class}}_3$-CA $+V = L[X]]$, by analogy with $\Pi^{1-\text{set}}_3$-CA of Simpson (1999) VII.3. This model would satisfy Σ^1_3-Uniformisation, from which Π^1_3-REC could be shown; this, together with Lemma 5 is what is needed for the proof to go through.

References

Beall, J., ed. (2003) *Liars and Heaps*, OUP, Oxford.

Boolos, G. (1984) "To Be is to Be a Value of a Variable (or to be Some Values of Some Variables)," *The Journal of Philosophy* 81, 430–49. Reprinted in Boolos (1998).

Boolos, G. (1985a) "Nominalist Platonism," *Philosophical Review* 94, 327–44. Reprinted in Boolos (1998).

Boolos, G. (1985b) "Reading the *Begriffsschrift*," *Mind* 94, 331–334. Reprinted in Boolos (1998).

Boolos, G. (1998) *Logic, Logic and Logic*, Harvard, Cambridge, Massachusetts.

Devlin, K. (1984) *Constructibility*, Perspectives in Mathematical Logic, Springer Verlag, Berlin, Heidelberg.

Field, H. (2003a) "A Revenge-Immune Solution to the Semantic Paradoxes," *Journal of Philosophical Logic* 32, 139–77.

Field, H. (2003b) "The Semantic Paradoxes and the Paradoxes of Vagueness." In Beall (2003).

Field, H. (2004) "The Consistency of the Naive Theory of Properties," *Philosophical Quarterly* 54, 78–104.

Hossack, K. (2000) "Plurals and Complexes," *The British Journal for the Philosophy of Science* 51:3, 411–443.

Kripke, S. (1975) "Outline of a Theory of Truth," *Journal of Philosophy* 72, 690–716.

Linnebo, Ø. (2003) "Plural Quantification Exposed," *Noûs* 37, 71–92.

McGee, V. (1997) "How We Learn Mathematical Language," *Philosophical Review* 106, 35–68.

McGee, V. (2000) "'Everything'." In Sher and Tieszen (2000).

Moschovakis, Y. (1974) *Elementary Induction on Abstract structures*, volume 77 of *Studies in Logic series*, North-Holland, Amsterdam.

Oliver, A., and T. Smiley (2001) "Strategies for a Logic of Plurals," *Philosophical Quarterly* 51, 289–306.

Parsons, C. (1990) "The Structuralist View of Mathematical Objects," *Synthese* 84, 303–346.

Quine, W. V. (1986) *Philosophy of Logic, Second Edition*, Harvard, Cambridge, Massachusetts.

Rayo, A. (2002) "Word and Objects," *Noûs* 36, 436–464.

Rayo, A. (forthcoming) "Beyond Plurals." In Rayo and Uzquiano (forthcoming).

Rayo, A., and G. Uzquiano (1999) "Toward a Theory of Second-Order Consequence," *The Notre Dame Journal of Formal Logic* 40, 315–325.

Rayo, A., and G. Uzquiano, eds. (forthcoming) *Absolute Generality*, Oxford University Press, Oxford.

Rayo, A., and S. Yablo (2001) "Nominalism Through De-Nominalization," *Noûs* 35:1, 74–92.

Resnik, M. (1988) “Second-Order Logic Still Wild,” *Journal of Philosophy* 85:2, 75–87.

Sher, G., and R. Tieszen, eds. (2000) *Between Logic and Intuition*, Cambridge University Press, New York and Cambridge.

Simpson, S. (1999) *Subsystems of second-order arithmetic*, Perspectives in Mathematical Logic, Springer Verlag, Berlin, Heidelberg.

Welch, P. (2003) “Ultimate truth *vis à vis* stable truth.” Submitted to Journal of Philosophical Logic.

Williamson, T. (2003) “Everything” 415–465. In Hawthorne, J. and D. Zimmerman, eds. *Philosophical Perspectives 17: Language and Philosophical Linguistics*, Blackwell, Oxford.