THE REVIEW OF SYMBOLIC LOGIC
Volume 4, Number 3, September 2011

COMMENTS ON MARTIN'S AND WELCH'S COMMENTS

HARTRY FIELD
Department of Philosophy, New York University

Again, my thanks to Tony Martin and Philip Welch for their careful reading of the book and their extremely interesting comments. Unfortunately there's only time to discuss a small fraction of what they say. I'll start with some themes common to the two of them.

1a. One such theme concerns the complexity of my proposed logic of the conditional. As Tony emphasizes, the internal Kripke theory of the minimal Kleene fixed point is simple and natural. (Tony's reformulation of Kripke in terms of infinitary logic is a nice way of seeing this.) In adding a conditional to this theory that continues to avoid paradox, it seems inevitable that this simplicity will be lost.

In fact, as both Tony and Philip emphasize, if you explain logical consequence in terms of the specific way that I propose for extending ground models, then the resulting logic of conditionals comes out highly unaxiomatizable. But there's no obvious reason for building all the details of that method of extension into the logical consequence relation. Instead, one could explain logical consequence in a way that made it correspond to some axiomatic system that my extension method shows to be compatible with the naive theory of truth; from that point of view, my method of extending ground models shows a way to save naive truth consistent with the chosen axiomatic logic, but also validating some nonlogical principles as well. Admittedly, it's likely to be somewhat arbitrary which axiomatic system counts as giving "the real logic" as opposed to the nonlogical principles that my method of extension validates, but I don't regard that as a problem: the boundaries of logic do seem to me partly arbitrary.

Indeed, the situation is familiar from other cases where natural semantic characterizations of consequence would make the logic unaxiomatizable, as with second-order logic or the logic of 'there are infinitely many'. There too, if we want the logic to be axiomatizable we must regard some of the principles (e.g., the analogs of set-theoretic statements not settled in ZFC) as not part of the logic properly so-called.

There is a further element in the case of my proposal for extending ground models that arguably is not present in second-order logic: some of the choices that go into my semantics do seem plainly arbitrary (e.g., starting the revision procedure with all conditionals getting value $\frac{1}{2}$); whereas *if you think that we have a determinate notion of set* (as opposed to a rather schematic notion that can be filled out in different ways by different choices as to, say, the size of the continuum), then you may think that second-order logic does not involve such arbitrariness. Basically I'd say that to the extent that a choice seems arbitrary, we shouldn't count it as built into the logic. Of course, this makes all the clearer that I haven't sharply demarcated what's part of the logic and what isn't.

The reason I'm not so worried about what counts as logic and what doesn't is that the point of my program is to show that we can accept the naive theory of truth (and satisfaction, and properties, etc.) *in a logic with a well-behaved conditional*, one that obeys

Received: May 6, 2010

© Association for Symbolic Logic, 2011

doi:10.1017/S1755020311000025

https://doi.org/10.1017/S1755020311000025 Published online by Cambridge University Press

principles that a conditional should obey. To show this, I show that we can accept such a naive theory *even with the upper bound on logics that my method of extending ground models provides* (and indeed that we could do so even on several variations of that upper bound, given by alteration of some of the plainly arbitrary details). If we can accept the naive theory of truth (and satisfaction etc.) even with the upper bound(s) of logic, we can accept it for whatever "real logic" we decide on that is less powerful than the upper bound (or than one of the upper bounds).

1b. I don't take the approach to the conditional in the book to be the last word on the subject: I'm exploring other approaches that share its most important characteristics but would yield a smoother logic. (The logics that share these characteristics are, essentially, the *G-logics* of Field, 2007; you can take the model theory in the book as demonstrating that there actually are G-logics.) For instance: part of the complexity in the official logic of the book is that the conditional is incomparable to the material conditional as ordinarily defined, $\neg A \vee B$. But while it's important to my project that the conditional be in some ways weaker than the material conditional as usually defined (so that $A \rightarrow A$ can hold generally, despite the fact that $\neg A \vee A$ doesn't), it isn't really important that it be in other ways stronger; that was sort of an accident of my chosen model theory. The complexity that results from it being in some ways stronger is easily eliminable: we could define a modified conditional $A \rightarrow_{new} B$ as $(A \rightarrow_{old} B) \vee (\neg A \vee B)$, and the modified one would satisfy my requirements on a G-logic as well as the old one did. (It's not hard to give a model-theoretic semantics for the new conditional directly, without relying on the old conditional.) Some laws involving the conditional hold in stronger versions when it's the new conditional than when it's the old: for example, $A \rightarrow_{new} (B \rightarrow_{new} A)$ comes out valid, whereas with the old conditional we merely have the rule $A \models B \rightarrow_{old} A$. Since the new conditional is defined from the old, the proof that the naive theory of truth, satisfaction, and so on is consistent in the old logic of the conditional extends to the new.

In retrospect, I think I should have used the modified conditional rather than the old, since the modified one leads to more of the usual principles of classical logic. There would have been a downside: when searching for counterexamples to principles that still don't come out valid, the counterexamples to principles involving the new conditional tend to be more complicated. (I guess that's to be expected, since the new conditional is so much closer to the classical!)[1]

1c. Philip raises another worry about complexity: the fact that my model involves a transfinite construction that "closes" only at a very large countable ordinal. He says that this bothers most mathematicians (he jokes about "weapons of mass construction"), and wonders why philosophers aren't more bothered. I'd like to hear more about why we should be bothered.

I think at least part of his worry connects up with Feferman's idea of reflective closure of theories, where the reflective closure tells you all of the implicit commitments of the theory. One way to "reflectively close" on a mathematical theory such as Peano arithmetic or Zermelo–Fraenkel set theory is to add a nonstratified truth or satisfaction predicate to it, expanding the schemas (induction, replacement, etc.) to formulas containing the new

[1] The approach just sketched can be combined with the variation on my conditional mentioned at the start of section 17.5 of my book, where I use the Lukasiewicz three-valued semantics in the revision process. Even by itself, that tends to yield results closer to classical, with the result that counterexamples are harder to find; it would do so even more when combined with the approach mentioned in the text.

https://doi.org/10.1017/S1755020311000025 Published online by Cambridge University Press

predicate. (Some small adjustment of the schemas will also be needed, if the truth or satisfaction predicate doesn't obey excluded middle: for example, in arithmetic induction must be presented as a rule rather than as an axiom.) I think Philip's worry, at least in part, is to the idea that the "reflective closure" of arithmetic should be so powerful that it would need to be described by an ordinal process of such length, a length far exceeding the Church–Kleene ordinal. But I'm not clear exactly why this is objectionable: is it *obvious* that reflection about truth shouldn't give rise to results in number theory that go far beyond what can be achieved predicatively? (Indeed, I have some inclination to think that it's an advantage if the addition of a truth or satisfaction predicate to a theory greatly expands the power of the theory. But I've been thinking about this more in the context of adding a satisfaction predicate to set theory rather than number theory, and it may be that the pressure against excessive power is stronger in the number-theoretic case than in the set-theoretic.)

This connects closely with Tony's remark that I say almost nothing in the book about the question of the deductive power of bivalent parts of the language. He's correct, and I agree that such questions are extremely interesting; I neglected them in the book only because they exceed my technical competence. Adding a satisfaction predicate to arithmetic allows you to simulate quite a bit of second-order (and of still higher order) arithmetic; so the question arises how the bivalent part of the second-order arithmetic that you get from my proposed satisfaction predicate relates to existing subtheories of second-order arithmetic.[2] Some information about this seems implicit in Tony's and Philip's remarks, but I'm not very clear about the overall upshot and would like to see a more explicit and thorough discussion.

2. Another topic that both Tony and Philip discuss is the relation of the model I propose to my view of reality (i.e., my view of what's really true). The model is described classically and has its quantifiers limited to ranging over a set; reality contains too much to fit into a set, and according to me is nonclassical in that excluded middle fails for some sentences containing 'true'. So how can the model illuminate reality? This is a question I discuss extensively in the book. In partial summary, I say that it's a serious error to think that a theory of truth such as Kripke's or mine serves to *define* 'true'. Instead, we must regard 'true' as a primitive, our understanding of which must be conveyed by principles we accept, such as Intersubstitutivity. In developing this viewpoint I make several claims:

One claim is that there is a distinction between a sentence *being true* and its *having a designated value*, and that model-theoretic constructions (like Kripke's and mine) *at most* define what it is for a sentence to have a designated value, not what it is for it to be true. (This is the issue of truth-value gaps: in both Kripke's model theory and mine, the Liar sentence is not designated, but I'm unwilling to infer that it isn't true.)

A second claim is that model-theoretic constructions like Kripke's and mine can't really define even what it is for a sentence to have a designated value, except for sentences in a language whose quantifiers are restricted to a set. (That this is so should be fairly uncontroversial, since these constructions all start from a set-model of the 'True'-free ground language.) So if we want to apply the construction to English, or to the language of set theory, then what is defined isn't 'has a designated value' but 'would come out as

[2] A direct comparison is complicated. Unlike any proper subsystem of full second-order arithmetic, my theory allows totally unrestricted comprehension. But while excluded middle will be completely unproblematic for formulas that are broadly predicative, it can't be assumed without restriction; and where we don't have it we have problems recovering extensionality.

https://doi.org/10.1017/S1755020311000025 Published online by Cambridge University Press

having a designated value if its quantifiers were restricted to domain M (a set)'. Call this *domain-relative designatedness*.

A third claim (which may not be quite explicit in the book) is that there's no obvious reason to think we even understand a notion of designatedness that is not domain-relative.

A fourth is that by contrast, our primary notion of truth is not in this way domain-relative: we understand perfectly well what it is for a sentence with unrestricted quantifiers to be true.

If all these claims are accepted, then what the model-theoretic constructions that Kripke and I give are very far from defining 'true'. Their interest, rather, is that they are model theories *for sentences that can contain 'true'*. The way that model theories illuminate is by illuminating the logics of the notions in the languages they deal with. So these model theories illuminate the combined logic of 'true' and the other logical connectives.

Either of the first two claims would suffice for my main point, that the model theories do not define truth. And Tony agrees with both of these claims, but he does suggest that I have overemphasized the second one; this may mean that he disagrees with my third claim. At any rate, what he says is that though we can't literally *define* a notion of designatedness that is not domain-relative, we can come *closer* to a definition of it if, instead of employing Kripke's inductive definition of the minimal fixed point, we use an alternative characterization in terms of what follows from the totality of "ground level facts" in a simple and intuitive infinitary logic (though one whose consequence relation is not literally definable). Is this enough to clarify a domain-independent notion of designatedness for the Kripke theory of minimal fixed points? I'm not sure. But at any rate, as Tony says, it doesn't apply to my own logic, or even to the Kripke logic for nonminimal fixed points. For Tony this is further reason to feel slightly dissatisfied by theories that extend beyond the theory of the minimal strong Kleene fixed point—even though he appreciates the reasons for wanting to do so.

I'm somewhat less clear about what Philip thinks of my four points. On a somewhat related matter, he thinks, with Hannes Leitgeb, that my program will only be successful in the end if we can give a model theory in nonclassical terms which allows for models that include domains too big to be a set. As I say in the book, I don't rule out the possibility of such a model theory, and think that some of what I do in the book gives suggestions for how it might go. But I don't see why my program for the paradoxes rests on the possibility of such a theory: that involves raising higher demands on my program than are usual for classical logic. For instance, in the case of classical logic the usual account of validity involves only set-sized models. In response, Philip might argue that that's satisfactory in the classical case only because of Kreisel's squeezing argument, which relies on the completeness of classical logic. But in response to this response, I say that even in absence of completeness the model theory can yield an upper bound on the consequence relation, which is all we need in order to know that the logic is one in which the naive theory of truth and satisfaction is consistent.

3. The end of Philip's paper has some interesting remarks about my hierarchy of determinateness operators, a topic that's related to my claims to have provided a theory in which no "revenge paradoxes" can be constructed. To review, the logic of conditionals allows one to construct a determinateness operator D, where DA is in a certain sense strictly stronger than A (though equivalent to A as regards sentences that obey excluded middle.) We can iterate D—getting DD, DDD, and so on—and using the truth predicate one can extend the iteration far into the transfinite. Given a transfinite iteration D^α of D, one can construct an extended Liar sentence Q_α which declares itself not

https://doi.org/10.1017/S1755020311000025 Published online by Cambridge University Press

D^α-true. Q_α is declared not to be $D^{\alpha+1}$-true in the theory. It would violate the principles of naive truth if Q_α were declared not to be D^α-true, but it never is. Clearly this rests on $D^{\alpha+1}$ always being strictly stronger than D^α, for any α for which the transfinite iteration D^α makes sense.[3]

One might still try to construct a revenge paradox, by defining a sentence as *super determinately true* if it is D^α-true *for every α for which the transfinite iteration D^α makes sense*. The book contains a lengthy discussion of why this fails, but one can get some sense of this by reflecting that *if* it worked, it would in effect make sense of a transfinite iteration D^λ for the first λ greater than all the α for which D^α makes sense, which of course is absurd. (There's a close connection to the König paradox of the smallest undefinable ordinal, about which more shortly.)

Philip sketches an intriguing alternative method of iterating determinateness operators, which allows for far longer iterations than the method I employed. He also makes some remarks that might seem to be at odds with what I say: in particular, "the internally defined determinateness hierarchy breaks down, not fuzzily, but precisely, at Δ_0." I don't think he takes this to be a deep conflict with what I say—the remarks that follow about the contrast between the internal and external viewpoints seem intended to express agreement with at least the fundamentals of my position in the book. But the remarks confused me at first, and to some extent still do. The reason is that while the remarks seem right for cases where we *have* an external standpoint (as for instance in the case he explicitly discusses, when we reason about the theory resulting from adding 'true' *to number theory, from the standpoint of set theory*), I was working in a context where (I claim) we do not have such an external standpoint in any serious sense. When the internal standpoint is all we have to go on, Philip's remarks about a precise point at which the hierarchy breaks down seem to me misleading.

In more detail, let L^+_{set} be the language of set theory with an added satisfaction predicate. The way I iterated the determinacy operator in the book was to use formulas in the language L^+_{set} that *define* ordinals in the sense of being uniquely satisfied by them. I defined D^α for variable α in such a way that as long as α is hereditarily definable in L^+_{set}, D^α defines a corresponding iteration of D. This is an "internal" definition of the iterations: it makes no reference to models, but instead uses the definability predicate of the theory, which is in turn defined in terms of satisfaction. Since the law of excluded middle doesn't apply in full generality to 'satisfies', there's no reason to expect it to apply in full generality to 'defines' either. And it doesn't; for if it did, we could argue that there is a smallest ordinal undefinable in L^+_{set}. That would generate König paradox: how can it be undefinable in L^+_{set}, since given naive satisfaction 'the smallest ordinal undefinable in L^+_{set}' would have to define it? The paradox disappears once we restrict excluded middle for 'defines'; and with that restricted, it becomes "fuzzy" precisely how far the method of iterating D based on definability is well behaved. (The breakdown does occur way above the Church–Kleene ordinal. But it must occur somewhere in the countable

[3] In the book I end up defining D^α for arbitrary ordinals, but when the ordinal is big enough the operator trivializes to the operator that yields a clear falsehood when applied to any sentence at all. So for such big α, the defined D^α doesn't really correspond to an iteration of D: it is simply a degenerate byproduct of a formula that does define iterations of D as long as it is well-behaved. When I speak of the iteration "making sense" I mean that the defining formula does correspond to an iteration of D.

https://doi.org/10.1017/S1755020311000025 Published online by Cambridge University Press

ordinals, because L^+_{set} is a countable language and hence can define only countably many ordinals.)

Philip's method of iterating the determinacy operator is different: it too is done "internally" using a truth predicate, but Philip explains it by reference to an external model-theoretic construction of the sort that I use in my book. Instead of using formulas that correspond to ordinals by being uniquely true of them, he uses sentences that correspond to ordinals by stabilizing at them in the model-theoretic construction.

That is an external description of how those sentences, and the iterations of D based on them, correspond to ordinals. It's clear what it means in the case he focuses on, where L_{num} is the language of number theory and L^+_{num} the result of adding a truth predicate to that. On his internal analog of this, we define within L^+_{num} a formula $P_<(x, y)$; and we prove, in set theory without a truth predicate, that it "model-theoretically defines" a well-ordering whose length corresponds to $\Delta_{0,num}$, where that is the ordinal at which my model-theoretic construction terminates when started from the standard model of the natural numbers. (The external model theory defines the ordering by settling which pairs of sentences A, B are such that $P_<(\langle A\rangle, \langle B\rangle)$ gets value 1.) We then get an iteration of D corresponding to any proper initial segment of the ordering. But as Philip explains, from within L^+_{num} one is in no position to assert excluded middle for $P_<$, so that there is no internally definable maximal hierarchy.

As Philip and Tony made clear to me at the American Philosophical Association session, the picture can be applied to models of set theory too.[4] If M is a model of ZFC, let L^+_M be the result of adding to the language of set theory a name for every member of the domain of M and a truth predicate for sentences in this language. (We could forego the names and add a satisfaction predicate instead of a truth predicate, but if I understand Philip correctly it has to be a satisfaction predicate that is restricted to the domain of M.) Then we define a formula $P_<(x, y)$ much as before. (Philip tells me that the precise definition will depend on M, at least in that it depends on which names we've added or which restriction we place on the satisfaction predicate.) And we prove as before that it "model-theoretically defines" a well-ordering whose length corresponds to $\Delta_{0,M}$, where that is the ordinal at which my model-theoretic construction terminates when started from ground model M.

A nice thing about Philip's method is that it yields iterations of D *far* longer than the ones I got: indeed when the starting model M is uncountable, we will get uncountable iterations. (Of course I could have gotten uncountable iterations too, by adding uncountably many names of ordinals to the language of set theory; but that would seem a cheap trick in that context, whereas it seems more natural in Philip's.)

But there seems to me a slight oddity in basing an iteration procedure on a specific model of set theory. This especially seems so when the set theory is the most comprehensive one we have available. Suppose that the classical part of the theory is ZFC plus all your favorite large cardinal axioms. Call its language L_{set}, though we should allow for the possibility that it includes classical predicates not in the usual set-theoretic language. Now expand the language as before by adding a naive satisfaction predicate (restricting excluded middle), getting language L^+_{set}. Given the comprehensiveness of the starting set theory, we have no useful way to "go outside" it, unless we have enough genius to come up with more powerful mathematical axioms (e.g., for large cardinals) that were hitherto unavailable. Until such time as we're in a position to do that, there's no way to argue that this whole theory has

[4] I'm extremely grateful to both of them for their help on that occasion, and to Philip for clearing up a number of further confusions later on.

https://doi.org/10.1017/S1755020311000025 Published online by Cambridge University Press

models at all, though to the extent we believe it consistent we believe it does; even less is there a way to argue that it has models *of any very natural sort*. We can of course reason hypothetically from its having a model M: in particular, we can use the model-theoretic construction of $\Delta_{0,M}$ from the hypothetical M to "externally" reason about the truth theory of the language L^+_M. But the internal theory we're interested in isn't for L^+_M, it's for L^+_{set} with its more extensive satisfaction predicate. Given the apparent dependence of the formula $P_<(x, y)$ used for L^+_M on the choice of M, I would assume that it is not possible to define a formula $P_<(x, y)$ appropriate to L^+_{set} rather than to the various L^+_M. But even if it is, I don't know exactly what the constraints on it would be, since we don't have a model-theoretic construction based on the whole universe of sets, or any analog of a Δ_0 in this context. In this context, I can see no clear sense to saying that there is a precise point at which the iteration procedure breaks down.

§1. Acknowledgement. This paper was presented at a meeting of the American Philosophical Association in New York in December 2009.

BIBLIOGRAPHY

Field, H. (2007). Solving the paradoxes, escaping revenge. In Beall, J. C., editor. *The Liar's Revenge*, Oxford: Oxford University Press, pp. 78–144.

DEPARTMENT OF PHILOSOPHY
NEW YORK UNIVERSITY

https://doi.org/10.1017/S1755020311000025 Published online by Cambridge University Press