New Grounds for Naïve Truth Theory

S. Yablo (MIT)

A theory is semantically closed if it contains, for each sentence $\underline{A}$ of the language in which it is framed, all biconditionals of the form $T[\underline{A}] \equiv \underline{A}$.[1] Tarski showed that no consistent 1st order theory with a good grip on its language's syntax could be semantically closed. This is because a theory with a good grip on its language's syntax will (by the diagonal lemma) contain for each formula $\varphi(\underline{x})$ a biconditional $\underline{E} \equiv \varphi[\underline{E}]$, hence in particular a biconditional $\underline{L} \equiv \neg T[\underline{L}]$. A theory with $\underline{L} \equiv \neg T[\underline{L}]$ cannot in consistency also contain $T[\underline{L}] \equiv \underline{L}$. But that is what semantic closure requires.

[1] Every sentence is assumed to have at least one name in the language. '$[\underline{A}]$' is schematic over names of $\underline{A}$.

There is no denying Tarski's result, but one can try to steer around it. The ways of steering around it correspond to a number of things that Tarski did not show.

∞ He did not show that a theory with a *poor* grip on its language's syntax could not be consistent and semantically closed. One way to avoid Tarski's result is to insist that all reference to sentences be by means of quotation names.[2] There is nothing to prevent a consistent theory from containing all biconditionals of the form T'*A*' ≡ *A*, which on the stated hypothesis is all the T-biconditionals there are. (See Gupta 1982.)

[2] I am oversimplifying, since paradoxes can also be fashioned out of syntactic predicates. See Gupta 1982 for details.

∞ He did not show that a consistent, syntactically resourceful theory could not be *weakly* closed in the sense of containing, not perhaps *all* T-biconditionals, but at least one such biconditional for each *A*. A second way to avoid Tarski's result is to construct the T-biconditional for *L* using a different name from the one the theory uses to affirm that *L* ≡ ¬T[*L*]. (See Skyrms 1984.)

∞ Tarski did not show that a consistent, syntactically resourceful theory could not be *partly* closed in the sense of containing, not perhaps all T-biconditionals for *every* sentence *A*, but all for sentences *A* of a particularly well-behaved type. A third way to avoid Tarski's result is to require T-biconditionals for "nice" sentences only, settling for approximations thereto when *A* is not nice. (See Feferman 1984.)

∞ He did not show that a consistent, syntactically resourceful theory could not be *quasi*-closed in the sense of containing all sentences of the form $T[\underline{A}] \; \# \; \underline{A}$, where # expresses a type of equivalence other than the type expressed by ≡. A fourth way to avoid Tarski's result is to add a non-classical conditional → to the language and include in your theory all instances of $(T[\underline{A}] \rightarrow \underline{A}) \wedge (\underline{A} \rightarrow T[\underline{A}])$, that is, $T[\underline{A}] \leftrightarrow \underline{A}$. (See Brady 1989.)

∞ Tarski did not show that a semantically closed theory could not be consistent by the lights of a suitably chosen *non*-classical logic. A fifth way to avoid Tarski's result is to concede classical inconsistency but maintain that this or that classical rule is in the present context invalid. (See Priest 1979.)

****

All of these strategies have been tried, separately and in combination. But the last two have been tried the least. Field's approach combines elements of both. He adds a non-classical connective → while at the same time de-classicalizing the logic of the other connectives (by rejecting excluded middle). This makes for a remarkably powerful package, as you can see from the list now provided of its

Top Ten Excellent Features

1. Verifies all sentences of the form $T[\underline{A}] \leftrightarrow \underline{A}$[3]
2. Makes $T[\underline{A}]$ and $\underline{A}$ substitutable "salva valutate"
3. Provides an explicit model, hence…
4. No possibility of hidden paradoxes

[3] Here and throughout, $\underline{C} \leftrightarrow \underline{D} =_{df} (\underline{C} \rightarrow \underline{D}) \wedge (\underline{D} \rightarrow \underline{C})$.

5. High degree of revenge-immunity, even though…
6. Vengeance-threatening notions are expressible
7. → has a natural semantics
8. → has a ⊃-like logic
9. → "becomes" ⊃ in bivalent contexts
10. Only theory to inspire a Top Ten list.[4]

These features will come in for further explanation below, but let me say now what is meant by feature 2. T[A] and A are substitutable "salva valutate" iff for all sentential contexts φ(…), →-contexts not excluded, φ(T[A]) agrees with φ(A) on whatever semantic values there happen to be.

Well, what semantic values do there happen to be? Field uses a three-valued scheme in which sentences are assigned either 1 (determinate truth, or something like it), 0 (determinate falsity), or

[4] OK, so I ran out of ideas after nine.

1/2 (indeterminate). Values are generated by a transfinite series P^{α} of Kripkean fixed points, each the Kripke-closure (see below) of a “seed” valuation S^{α} that assigns values only to conditionals, defined here as statements whose main connective is →. Each P^{β} gives clues to the proper interpretation of →, clues that guide the construction of $S^{\beta +1}$, which forms the basis for $P^{\beta+1}$. This results in the back-and-forth process shown in Figure 1: [5]

[5] Yablo 1985 used a series of Kripkean fixed points to interpret the truth predicate, which was seen as nonmonotonic. Mention was made at the end of extending the method to "Lukasciewicz implication" (?), but I seem to remember having no idea what I was talking about.

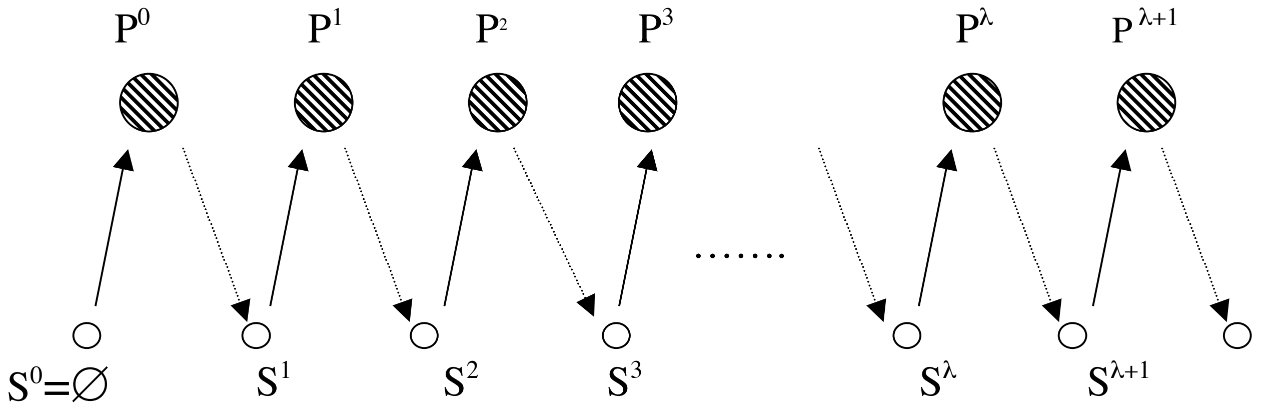


Figure 1

Three things need to be explained in this diagram. (a) How is S^α built up into P^α? (b) How is S^α calculated on the basis of earlier P^βs? And (c) how does the process determine semantic values for sentences? The answer to (a) comes from Kripke. The answer to (c) comes from Herzberger and Gupta. The answer to (b) is new with Field and the key to his construction.

To begin with (a), we can think of valuations (e.g., S^{α} and P^{α}) as sets of ordered pairs $<\underline{A}, v>$, where $\underline{A}$ is a sentence and v is a truth-value, either 0 or 1. A valuation V is *Kleene* (strictly, *Kleene over M*, for M a model of the base language) iff

(1) $<F\underline{a},1> \in V$ iff $M(\underline{a}) \in M(F)$[6]

(2) $<F\underline{a},0> \in V$ iff $M(\underline{a}) \notin M(F)$

(3) $<\neg\underline{A},1> \in V$ iff $<\underline{A},0> \in V$

(4) $<\neg\underline{A},0> \in V$ iff $<\underline{A},1> \in V$

(5) $<\underline{A}v\underline{B},1> \in V$ iff $<\underline{A},1> \in V$ or $<\underline{B},1> \in V$

(6) $<\underline{A}v\underline{B},0> \in V$ iff $<\underline{A},0> \in V$ and $<\underline{B},0> \in V$

(7) $<\forall\underline{x}\ G\underline{x},1> \in V$ iff $<G\underline{a},1> \in V$ for each name $\underline{a}$[7]

(8) $<\forall\underline{x}\ G\underline{x},0> \in V$ iff $<G\underline{a},0> \in V$ for some name $\underline{a}$

[6] Assume for simplicity that all predicates are monadic.

[7] Assume that every object has a name.

V is *Kripke* iff

(9) $<T[\underline{A}],1> \in V$ iff $<\underline{A},1> \in V$

(10) $<T[\underline{A}],0> \in V$ iff $<\underline{A},0> \in V$

A *fixed point* – of Kripke's jump operator, but let that be understood -- is a valuation that is both Kleene and Kripke. Provided a valuation V is *sound* in the sense of satisfying the left-to-right directions of (1)-(10), it can be built up into a fixed point V* by closing under the right-to-left directions of (1)-(10). Note that S^α is automatically sound because it assigns truth-values only to conditionals, and there are no conditionals on the left sides of (1)-(10). The relation between P^α and S^α is simply that $P^\alpha = S^{\alpha}*$.

Now let's consider (b) how S^α is calculated on the basis of earlier P^βs. Given our treatment of valuations as (not necessarily single-valued) relations between sentences and truth-values, *A*'s semantic value in V is best understood as $\{v \mid <\underline{A},v> \in V\}$. Thus, limiting ourselves for now to consistent valuations, the semantic values are {1}, {0}, and {}. These values are ordered in the obvious way: {1} > {} > {0}. (Sometimes we use Field's notation and write {1} as 1, {0} as 0, and {} as 1/2; then the ordering is by numerical size.) If α is a successor ordinal $\beta+1$, then S^α is

$$\{<\underline{A}\rightarrow\underline{B},1>\mid P^\beta(\underline{A}) \leq P^\beta(\underline{B})\} \cup \{<\underline{A}\rightarrow\underline{B},0>\mid P^\beta(\underline{A}) > P^\beta(\underline{B})\}.$$

If α is a limit ordinal, then S^α is

$$\{<\underline{A}\rightarrow\underline{B},1>\mid \exists\gamma<\alpha\ \forall\beta\in[\gamma,\alpha)\ \ P^\beta(\underline{A}) \leq P^\beta(\underline{B})\}$$
$$\cup\{<\underline{A}\rightarrow\underline{B},0>\mid \exists\gamma<\alpha\ \forall\beta\in[\gamma,\alpha)\ \ P^\beta(\underline{A}) > P^\beta(\underline{B})\}.$$

The S^αs and P^αs will as Field says "oscillate wildly" (Figure 1 is in that respect misleading), but the hope is that deserving sentences will eventually stabilize on their deserved classical value (1 or 0), while defective sentences will stabilize at 1/2 or never stabilize at all. Which brings us to (c), the assignment of ultimate values. $\|A\|$ is $\lim_{\beta} P^{\beta}(\underline{A})$ if the limit exists, otherwise 1/2. That is, $\|\underline{A}\| = 1$ (0) if $P^{\beta}(\underline{A})$ is eventually always 1 (0), and $\|A\| = 1/2$ if $P^{\beta}(\underline{A})$ is eventually always 1/2 or not eventually always anything.

****

The main and most exciting claim Field makes on behalf of his semantics is that it validates the "naïve theory of truth":

(i) $\|T[\underline{A}] \leftrightarrow \underline{A}\| = 1$ unrestrictedly, and

(ii) $\|\ldots T[\underline{A}]\ldots\| = \|\ldots\underline{A}\ldots\|$ unrestrictedly.

Property (i) and most of property (ii) are immediate from the fact that each P^{β} is a Kripkean fixed point. For it is a feature of every fixed point that $P^{\beta}(T[\underline{A}]) = P^{\beta}(\underline{A})$, and that $T[\underline{A}]$ and $\underline{A}$ are freely substitutable in other-than-conditional contexts. Substitutivity within the scope of → is proved by induction on the complexity of $\underline{A}$. For instance, $||\underline{B} \rightarrow \underline{C}|| = 1$ iff $P^{\beta}(\underline{B})$ is eventually always $\leq P^{\beta}(\underline{C})$ iff $P^{\beta}(T[\underline{B}])$ is eventually always $\leq P^{\beta}(\underline{C})$ iff $||T[\underline{B}] \rightarrow \underline{C}|| = 1$. Field proves in fact that ||…|| is one of the P^{β}s and so a full-fledged Kripkean fixed point.

Note that to call ||…|| a Kripkean fixed point is not to say that it is a fixed point of Field's operator, the operator taking P^{β} to $P^{\beta+1}$. Recall that a Kripkean fixed point needs only to satisfy (1)-(10) above. The closure S* of any seed set S, obtained by slapping 1's and 0's on conditionals however one likes, does that much. To be a Fieldian fixed point, V must also satisfy

(11) $<\underline{A} \rightarrow \underline{B}, 1> \in V$ if $V(\underline{A}) \leq V(\underline{B})$

(12) $<\underline{A} \rightarrow \underline{B}, 0> \in V$ if $V(\underline{A}) > V(\underline{B})$.

This means that V must be S* not for any old S but the particular one S_V that V induces:

$$\{<\underline{A}\rightarrow\underline{B},1>|\ V(\underline{A}) \leq V(\underline{B})\} \cup \{<\underline{A}\rightarrow\underline{B},0>|\ V(\underline{A}) > V(\underline{B})\}.$$

But if the language contains a Curry sentence – a $\underline{K}$ identical or equivalent to $T[\underline{K}]\rightarrow 0=1$ – then no V can do this. Either $V(T[\underline{K}]) \leq V(0=1)$ or $V(T[\underline{K}]) > V(0=1)$. If the latter, then by (12), $V(T[\underline{K}]\rightarrow 0=1) = V(\underline{K}) = 0$, whence (by (10)), $V(T[\underline{K}]) =0 \leq V(0=1)$ after all. If the former, then by (11), $V(\underline{K}) = 1$, whence (by (9)) $V(T[\underline{K}]) = 1$, contradicting our assumption that $V(T[\underline{K}]) \leq V(0=1)$.

How much should it bother us that $\|....\|$ is not a Fieldian fixed point? What is nice about (11) and (12) is that they equip →-sentences with intuitive and comprehensible truth-conditions. By the same token, it should be no great cause for alarm if the “real” truth-conditions don't take precisely the (11)-(12) form, provided that intuitive comprehensibility is not sacrificed,

Indeed (11) and (12) might be considered problematic, since they prevent $\underline{A}\rightarrow\underline{B}$ from assuming the value 1/2, thus obliterating the distinction between conditionals whose antecedents are *much* truer than their consequents $(1\rightarrow 0)$ and ones whose antecedents are only a little truer $(1 \rightarrow 1/2$, or $1/2 \rightarrow 0)$. Better from this perspective would be an interpretation along the lines of

(11') $<\underline{A}\rightarrow\underline{B},1> \in V$ if $V(\underline{A}) \leq V(\underline{B})$

(12') $<\underline{A}\rightarrow\underline{B}, V(\underline{A})-V(\underline{B})>$ if $V(\underline{A}) > V(\underline{B})$

(the 3-valued Lukasciewicz conditional). However, (11') and (12') are no more within reach than (11) and (12). Consider $\underline{K}$ as above $(T[\underline{K}]\rightarrow 0=1)$ and $\underline{K}'$ which says that $0=0\rightarrow\underline{K}$. At odd stages $\underline{K}$ and $\underline{K}'$ are 1 and 0; at even non-limit stages they are 0 and 1. Since

$\|\underline{K}\|$ and $\|\underline{K}'\|$ are 1/2, $\|\underline{K}\rightarrow\underline{K}'\|$ should according to (11') be 1. But it can't be 1, for at odd stages $\underline{K}$ has a higher value than $\underline{K}'$, making $\underline{K}\rightarrow\underline{K}'$ 0 at even successor stages. $\|\underline{K}\rightarrow\underline{K}'\|$ is in fact 1/2. That some 1/2→1/2 conditionals, like $\underline{K}\rightarrow\underline{K}$, are 1, while others, like $\underline{K}\rightarrow\underline{K}'$ are 1/2, shows that → is not value-functional. Therefore nothing like (11)-(12) and (11')-(12') can possibly work.[8]

****

Of course Field is under no illusions about this. He is not for a moment suggesting that → is a standard issue extensional connective. Still, it is natural to wonder how in that case → is to be understood. I am not sure that a commonsensical explanation of →'s meaning is possible. What we can do is try to place it on the map between two more familiar sorts of connective.

[8] See Field 2003a for a type of semantic value with respect to which → is compositional.

One is the extensional connective $\boldsymbol{\rightarrow}$ defined by (11') and (12'), the Lukasciewicz conditional. A look at the truth tables shows that $\rightarrow$ is stronger than $\boldsymbol{\rightarrow}$, in that $\|A\rightarrow B\|$ is never greater than $\|A\boldsymbol{\rightarrow}B\|$ and occasionally less, for instance $\|K\rightarrow K'\| = 1/2$ while $\|K\boldsymbol{\rightarrow}K'\| = 1$. This raises the question, in what does the additional strength consist? What more is asserted by $A\rightarrow B$ than $A\boldsymbol{\rightarrow}B$?

Necessitating a claim is the obvious way to strengthen it, so one natural conjecture is that $A\rightarrow B$ is $\Box(A\boldsymbol{\rightarrow}B)$ for some suitable modal operator $\Box$. But $\rightarrow$ does not appear to pack much modal punch. If it did, then one would expect $A\rightarrow B$ *not* to hold unless A somehow necessitated B, and $B\rightarrow A$ not to hold unless B necessitated A. Since "most" pairs of sentences necessitate in neither direction, one would expect $\|(A\rightarrow B)\vee(B\rightarrow A)\|$ to not often be 1. But in fact it is *always* 1 except under rather special conditions. For $\|(A\rightarrow B)\vee(B\rightarrow A)\|$ not to be 1, we need first that neither $\underline{A}$ nor $\underline{B}$ stabilizes at a classical value, and second that at least one of $\underline{A}$, $\underline{B}$ does not stabilize at all.[9] If the first condition is met but not the second, as for instance when $\underline{A}$ and $\underline{B}$ are the Truthteller and the Liar, then $\underline{A}\rightarrow\underline{B}$ and $\underline{B}\rightarrow\underline{A}$ are *both* true. This suggests that $\underline{C}\rightarrow\underline{D}$, although stronger than $\underline{C}\boldsymbol{\Rightarrow}\underline{D}$, is not as strong as $\square(\underline{C}\boldsymbol{\Rightarrow}\underline{D})$.

What does it matter, one might say, whether → has an antecedently comprehensible meaning? A conditional that gets the job done in other respects is one we will learn how to use, and learning the use will teach us the meaning. This response is fair enough in principle. But it assumes that → does get the job done in other respects, and that a certain semantic obscurity is just the price that has to be paid. This may be right in the end. But it seems to me that →'s performance in other respects is not beyond

[9] These conditions are necessary for a value other than 1, but not sufficient.

criticism, and that attending to the criticism makes →'s meaning less obscure rather than more.

****

When we speak of getting the job done in other respects, we should distinguish Job 1 -- verifying the T-biconditionals – from Jobs 2, 3, 4, etc. -- assigning appropriate values to <u>other</u> sentences (sentences not of the form T[<u>A</u>]↔<u>A</u>). When it comes to the evaluation of other sentences, I see three areas where → could stand to be tweaked. The first is that → should be less arbitrary. <u>A</u>→<u>B</u> should not be assigned 1 when there is an equally good case for assigning it 0. The second is that → should be more grounded. Assignments should be made on some prior and independent basis, and the dependency chains should eventually bottom out. The third is that → should be stricter. If <u>A</u> contradicts <u>B</u>, for instance, we want not only that <u>A</u>→¬<u>B</u> is 1, but also that <u>A</u>→<u>B</u> is not 1.

Arbitrariness

One objection people have raised to counting *E* (the Truthteller) true is that one could just as easily consider it false. Construed as true, it verifies itself, construed as false, it falsifies itself, and there is no reason to prefer one scenario over the other.[10]   To treat the Truthteller as true, pure and simple, strikes us as arbitrary, which in a truth-value assignment seems a Bad Thing.

[10] I myself think that there is a reason and that the Truth-Teller is false (Yablo 1985, 1993). But I am not aware of having convinced anyone of this, so I treat the Truth-Teller as semantically underdetermined. A better example from my perspective would be *J* = '*S* is false', *S* = '*J* is false.' That really is underdetermined.

By the Conditional Truthteller, let's mean a sentence $\underline{F}$ to the effect that $(\underline{A}\rightarrow\underline{A})\rightarrow T[\underline{F}]$.[11] $\underline{F}$ describes itself as true-if-$\underline{A}\rightarrow\underline{A}$. Because $\underline{A}\rightarrow\underline{A}$ is a tautology, this should be the same as $\underline{F}$ describing itself simply as true. The hypothesis that $\underline{F}$ is true is self-justifying, and likewise the hypothesis that $\underline{F}$ is false. Since there is no reason to prefer one outcome to the other, to consider $\underline{F}$ true, pure and simple, seems unacceptably arbitrary.

However, that is the value the semantics assigns. Recall that at stage 0, all conditionals are 1/2, including $\underline{A}\rightarrow\underline{A}$ and $(\underline{A}\rightarrow\underline{A})\rightarrow T[\underline{F}]$ (= $\underline{F}$). That $\underline{F}$'s antecedent and consequent have the same value at stage 0 makes $\underline{F}$ 1 at stage 1. That $\underline{F}$'s consequent is 1 at stage 1 makes $\underline{F}$ 1 at stage 2, and so on indefinitely. "I am true if $\underline{A}\rightarrow\underline{A}$" has therefore an ultimate value of 1.[12]

[11] $\underline{A}$ can be any sentence you like.

[12] Another oddity is that $\|\underline{F}\|$ becomes 0 if in place of $\underline{A}\rightarrow\underline{A}$ we put 0=0. .

Groundedness

A second reason not to call the Truthteller true is that truth attributions should be based on prior facts, facts already in place before they are made. I call this a different reason because there are cases where a non-arbitrary assignment seems objectionable simply because there is no prior and independent basis for it.

Consider an infinite line-up of people each saying "I am speaking truly ⊃ the person behind me is speaking truly" ($\underline{S}_{\underline{n}}$ is T[$\underline{S}_{\underline{n}}$] ⊃ T[$\underline{S}_{\underline{n}+1}$]). The only classically defensible assignment is 11111….., since if any $S_{\underline{n}}$ were false it would have a false antecedent, making it true after all. Still we are hard put to regard any of these sentences as true. There is no basis for calling $\underline{S}_{\underline{n}}$ true that does not run through assignments to later $\underline{S}_{\underline{m}}$s for which the same problem arises.

Suppose now that the people in our infinite line decide to use arrows instead of horseshoes ($\underline{S}_{\underline{n}}$ is $T[\underline{S}_{\underline{n}}]\rightarrow T[\underline{S}_{\underline{n+1}}]$). There is no more reason to call the sentences true than before, but this is the assignment that the semantics makes. Each $S_{\underline{n}}$ is 1/2 at stage 0, so 1 at stage 1, and then we never look back. Such an outcome may not be arbitrary, but it does seem ungrounded. To suppose that $\underline{S}_{\underline{n}}$ is true is to suppose it has a true antecedent. But then its truth is owing to the truth of its consequent $T[\underline{S}_{\underline{n+1}}]$, with the buck being passed on forever down the line.

Strictness

If a thing is red, then it is not orange. This seems true not only of determinately red and determinately orange objects but also objects on the red/orange border. The incompatibility of red with orange is not limited to the clear cases on either side. Similarly, if a sentence is true, then it not false. This applies not only to determinately true and determinately false sentences but also to sentences on the border.[13] The incompatibility of true with false is not limited to the clear cases on either side. Facts like these ought to be expressible in the language. (This is more or less the problem of penumbral connections.)

Suppose that Jones, assured by her instructor that the Truthteller is true, asserts T'*E*', while Smith, assured by his instructor that it is false, asserts T'¬*E*'. It seems that Jones and Smith are saying incompatible things. If Smith is right in calling *E* false, then Jones is wrong to call it true. Similarly if Jones and Smith call each

[13] Does it apply to *all* sentences on the border? Some might think it false of the Liar that if it is true, then it is not false. I suspect it *seems* false only because of a felt tension with 'if the Liar is true then it is false,' and that we do best to regard both sentences as true. No such problems arise for the examples in the text.

other liars: $\underline{J} = \neg T[\underline{S}]$ and $\underline{S} = \neg T[\underline{J}]$.[14] They cannot both be right. If Jones's statement $\underline{J}$ is true, then Smith's statement $\underline{S}$ is false, and vice versa.

It would seem natural to try to register the incompatibility of T'$\underline{E}$' with T'$\neg\underline{E}$' by saying that T'$\underline{E}$' → $\neg$T'$\neg\underline{E}$'. This sentence is certainly true for Field. But does it express the incompatibility of T'$\underline{E}$' with T'$\neg\underline{E}$'? I would say not, since ||T'$\underline{E}$' → T'$\neg\underline{E}$'|| is also 1, and far from being incompatible with $\neg$T'$\neg\underline{E}$', T'$\underline{E}$' would seem to imply it. A similar situation arises if one tries to express the incompatibility of $\underline{J}$ with $\underline{S}$ by saying that $\underline{J} \rightarrow \neg\underline{S}$. This has ultimate value 1 as desired. However $\underline{J} \rightarrow \underline{S}$ also has ultimate value 1, even though it is precisely *not* the case that $\underline{J}$ is incompatible with $\neg\underline{S}$ ($= \neg\neg T[\underline{J}]$).[15]

[14] I assume that $\underline{S}$ and $\underline{J}$ are {} (in Field's notation 1/2).

[15] The last three paragraphs take issue with a claim that Field has since dropped. He now says that the contrariety of red with orange

consists in the fact that $\Box\forall\underline{x}$ ($\underline{x}$ is red $\rightarrow$ $\underline{x}$ is not orange), where $\Box$ is conceptual necessity. (??p. 20??). A couple of things still worry me, however. First, Field uses the unnecessitated conditional to express a related notion: an object $\underline{x}$ is red-to-orange iff $\underline{x}$ isn't red $\rightarrow$ $\underline{x}$ is orange. However if this conditional really expressed that $\underline{x}$ was red-to-orange, one would expect it not to be true that $\underline{x}$ isn't red $\rightarrow$ $\underline{x}$ is big, for $\underline{x}$ is not red-to-big. And it looks like it will be true if $\underline{x}$ is borderline big, or at least that is the natural assumption given the almost-extensionality of $\rightarrow$. A different explanation of red-to-orange avoids this problem: $\Box\forall\underline{y}$ ($\underline{y}$ is indiscernible in color from $\underline{x}$ $\rightarrow$ ($\underline{y}$ is not red $\rightarrow$ $\underline{y}$ is orange)). This will be false if "big" is substituted since we can choose $\underline{y}$ to be small. Second, suppose with Field that P is contrary to Q iff $\Box\forall\underline{A}\,(P[\underline{A}] \rightarrow \neg Q[\underline{A}])$. "True" and "untrue if 0=0" are intuitively speaking contraries, so we should have $\Box\forall\underline{A}\,(T[\underline{A}] \rightarrow \neg(0=0 \rightarrow \neg T[\underline{A}]))$. But $\forall\underline{A}\,(T[\underline{A}] \rightarrow \neg(0=0 \rightarrow \neg T[\underline{A}]))$ is not true on Field's semantics, since it does

A thing is orange if and only if it is redder than yellow things and yellower than red ones. A sentence A is true iff A. Both claims seem correct for definite and borderline cases alike. The equivalence of orange with redder-than-yellow-and-vice-versa is not limited to the clear cases on either side, and likewise the equivalence of T[A] with A. Our theory of these matters should somehow register these equivalences.

Kripke's theory comes close to registering the T-equivalence, in that each fixed point treats the two sides alike. The trouble is that this is for Kripke an essentially metalinguistic observation. He lacks an object language connective # with the property that $\|B\#C\|$

not hold for the Curry sentence K. ($\|T[K] \rightarrow \neg(0=0 \rightarrow \neg T[K])\| = 1/2$.)

$= 1$ iff $Q(B) = Q(C)$ for all fixed points Q.[16] This is because his connectives are monotonic, and no monotonic connective can permit the combination of $\|B \# C\| = 1$ with $\|B\| = \|C\| = 1/2$.

Now Field's → is not monotonic, so one might hope it would allow the equivalence of T[A] with A finally to be stated. This requires not just that $\|T[A]\leftrightarrow A\|=1$, but also that $\|T[A]\leftrightarrow\neg A\|$ not be 1 (leaving aside cases, like the Liar, where the second assignment is forced on us by the first). . However this is not what we find, since $\|T[G]\leftrightarrow\neg G\| = 1$ for every G that stabilizes at 1/2, including the Truthteller and the Tautologyteller ("this sentence is either true or not true"). A related situation arises with sentences A that do not stabilize at 1/2. Notice first that if A' is $0=0\rightarrow A$, then A and A' are intuitively speaking equivalent. One might hope then that $\|T[A]\leftrightarrow A'\| = 1$ and $\|T[A]\leftrightarrow\neg A'\| = 0$. In fact the reverse is true

[00] Kripke 1975 does discuss the possibility of adding a modal operator.

for certain choices of $\underline{A}$: $||T[\underline{A}]\leftrightarrow\underline{A}'|| = 0$ and $||T[\underline{A}]\leftrightarrow\neg\underline{A}'|| = 1$. (This happens for instance if $\underline{A}$ is a Curry sentence.) To be sure, when $\underline{A}$ has ultimate value 1 or 0, biconditionals framed with → behave as they should: $||T[\underline{A}] \leftrightarrow \underline{A}|| = 1$, $||T[\underline{A}] \leftrightarrow \neg\underline{A}|| = 0$. But that much we get already from the material biconditional ≡. The advantage of → is supposed to lie with its treatment of the $||\underline{A}|| = 1/2$ case.

****

A number of strategies might be tried to address these issues (I don't call them "problems" since my reactions may be idiosyncratic). The ones I will look at exploit the fact that Kripke's minimal fixed point P is, as the name implies, only one of a *class* of valuations all satisfying conditions (1)-(10) listed above. The idea very roughly will be to give a possible worlds semantics for →, with these valuations playing the role of worlds. I will first sketch a construction in the style of Field, building on Herzberger and Gupta; it helps with some of the examples just discussed but does not greatly clarify the truth-conditions of $\underline{A}\rightarrow\underline{B}$. This will be followed by a construction in the style of Kripke (and Yablo 1985), which addresses all of the examples and makes the truth-conditions relatively straightforward.

Field-style Possible Worlds Semantics

Call a fixed point "categorical" if it leaves →-statement unevaluated. The Field-style construction takes off from (i) the minimal fixed point P and (ii) the set **Q** of all categorical fixed points. Because no consistency requirement is imposed – $Q \in \mathbf{Q}$ can contain neither, either, or both of $<\underline{A},1>$ and $<\underline{A},0>$. -- the members of **Q** form a lattice under the operations

$$Q_1 \vee Q_2 = (Q_1 \cup Q_2)^*$$

$$Q_1 \wedge Q_2 = (Q_1 \cap Q_2)^*.$$

Each $Q \in \mathbf{Q}$ has a dual that turns Q's gaps into gluts and vice versa, leaving other sentences unchanged. In particular $\mathbf{Q}$ contains a (unique) maximal fixed point obtained by taking the dual of P. (See Woodruff 1984.)

The definition of stages is by a simultaneous induction starting from $\mathbf{Q} = \{Q_{\underline{i}} \mid \underline{i} \in \underline{I}\}$. Each $Q_{\underline{k}}$ is the starting point $Q_{\underline{k}}^{0}$ of a sequence $<Q_{\underline{k}}^{\alpha}>$, with $Q_{\underline{k}}^{\beta+1}$ determined by $\mathbf{Q}^{\beta} = \{Q_{\underline{i}}^{\beta} \mid \underline{i} \in \underline{I}\}$. The key to the construction is that $\underline{A} \rightarrow \underline{B}$ is true in $Q_{\underline{k}}^{\beta+1}$ iff $\underline{B}$ is valued as highly as $\underline{A}$ by all the members of $\mathbf{Q}^{\beta.}$ By "valued as highly as" I mean

$$Q(\underline{A}) \leq Q(\underline{B}) =_{df}$$

$$\langle \underline{A},1 \rangle \in Q \text{ only if } \langle \underline{B},1 \rangle \in Q, \text{ and}$$

$<\underline{B},0> \in Q$ only if $<\underline{A},0> \in Q$.[17]

If $Q(\underline{C})$ is considered {}, {1}, {0}, or {1,0} according as Q contains neither of $<\underline{C},1>$, $<\underline{C},0>$, the first only, the second only, or both, then this corresponds to the ordering that has $\{0\} <$ all other values, $\{1\} >$ all other values (and that's all).. Q is said to *validate* $\underline{A}\rightarrow\underline{B}$ iff $Q(\underline{A})\leq Q(\underline{B})$.

Suppose we have $\mathbf{Q}^{\alpha} = \{Q_{\underline{i}}^{\alpha} \mid \underline{i} \in \underline{I}\}$ in hand. Then $\mathbf{Q}^{\alpha+1}$ is the set of all $Q_{i}^{\alpha+1}$ $(\underline{i} \in \underline{I})$, where

$$Q_{\underline{k}}^{\alpha+1} =$$

$$(\{<\underline{A}\rightarrow\underline{B}, 1> \mid \text{every } Q_{\underline{i}}^{\alpha} \supseteq Q_{\underline{k}}^{\alpha} \text{ validates } \underline{A}\rightarrow\underline{B}\})\}$$

$$\cup \{<\underline{A}\rightarrow\underline{B},0> \mid \text{not every } Q_{\underline{i}}^{\alpha} \supseteq Q_{\underline{k}}^{\alpha} \text{ validates } \underline{A}\rightarrow\underline{B}\})^{*}.$$

[17] $Q(\underline{C}) = Q(\underline{D})$ means that $<\underline{C}, v> \in Q$ iff $<\underline{D},v> \in Q$.

$\underline{A}\to\underline{B}$ is true in $Q_{\underline{k}}^{\lambda}$ iff as γ approaches λ, it is eventually always the case that every $Q_{\underline{i}}^{\gamma} \supseteq Q_{\underline{k}}^{\gamma}$ validates $\underline{A}\to\underline{B}$.[18]

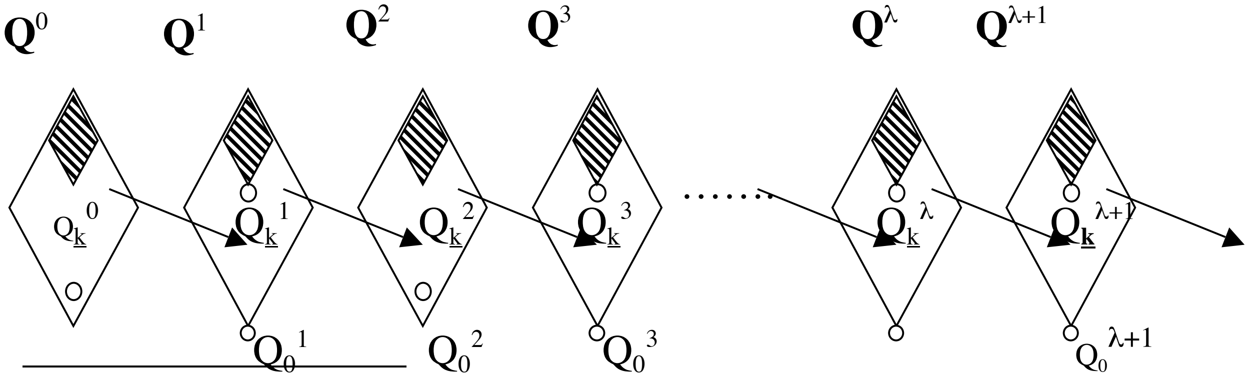
$Q_{\underline{k}}^{\lambda} =$

$(\{<\underline{A}\to\underline{B}, 1> \mid \exists\beta<\lambda\ \forall\gamma\in[\beta, \lambda)$ each $Q_{\underline{i}}^{\alpha} \supseteq Q_{\underline{k}}^{\alpha}$ validates $\underline{A}\to\underline{B}\}$

$\cup\ \{<\underline{A}\to\underline{B}, 0> \mid \exists\beta<\lambda\ \forall\gamma\in[\beta, \lambda)$ no $Q_{\underline{i}}^{\alpha} \supseteq Q_{\underline{k}}^{\alpha}$ validates $\underline{A}\to\underline{B}\})^*$.

This results in the back-and-forth process shown in

Figure 2:



[18] This looks a little different from the successor clause but is in keeping with Field..

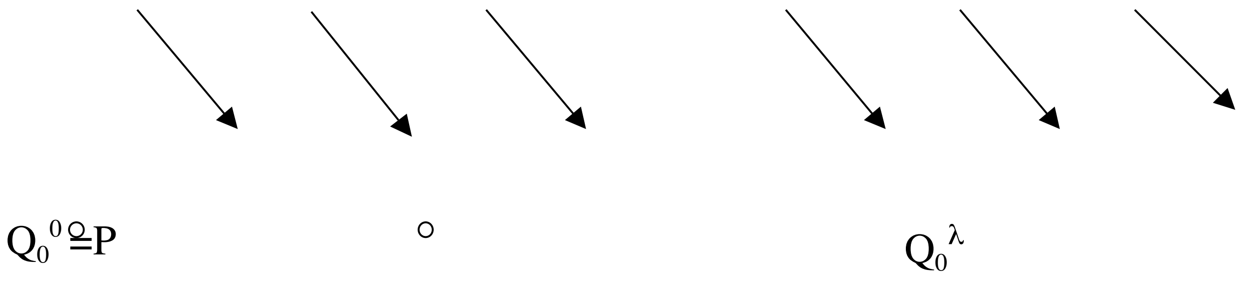


Figure 2

Now the definition of ultimate values. For each $Q_{\underline{k}}$ in **Q**, let $Q_{\underline{k}}^{\infty}$ be $\mathrm{liminf}_{\beta} Q_{\underline{k}}^{\beta,}$ that is

$$Q_{\underline{k}}^{\infty} = \{<\underline{C},v> \mid <\underline{C},v>\in Q_i^{\beta} \text{ for all large enough } \beta\}.$$

If Q_0 is the minimal fixed point, then $\underline{A}$'s ultimate value is the value it is assigned by Q_0^{∞}. Consider for instance "if $\underline{J}$, then not $\underline{S}$", where again $\underline{J}$ is $\neg T[\underline{S}]$ and $\underline{S}$ is $\neg T[\underline{J}]$. This receives value 1 because $Q(\underline{J}) \leq Q(\neg\underline{S})$ $(= Q(\neg\neg T[\underline{J}]))$ for all fixed points Q. "If $\underline{J}$, then $\underline{S}$" is not assigned 1 because it is not validated by consistent Qs assigning 1 to $\underline{J}$, and such Qs can be found in every $\mathbf{Q}^{\beta}$. (Recall that "if $\underline{J}$ then $\underline{S}$" is 1 on the original Field semantics.)

I do not want to dwell too long on this system, so let me just note a few basic properties. That $Q_{\underline{k}}^{\alpha}$ is always a Kripkean fixed point means that each Q_0^{α} validates both directions of $T[\underline{A}] \leftrightarrow \underline{A}$, so $Q_0^{\infty}(T[\underline{A}] \leftrightarrow \underline{A}) = 1$. Substitutivity is harder to prove, but it holds, so we have the full naïve theory of truth. → is stricter than it was than on the original Field semantics, which helps with penumbral connections (as just observed in effect with $\underline{J}$ and $\underline{S}$). But the issues of arbitrariness and groundedness remain. It is unfortunately <u>not</u> the case that $P^{\infty}(\underline{A} \rightarrow \underline{B}) = 1$ iff $\forall\ Q_{\underline{k}}^{\infty} \sqsupseteq Q_0^{\infty}\ \ Q_{\underline{k}}(\underline{A}) \leq Q_{\underline{k}}(\underline{B})$.[19] The new truth-conditions are as obscure as the old ones.

[19] The right hand side holds for $\underline{K}$, $0=0 \rightarrow \underline{K}$ but $P^{\infty}(\underline{K} \rightarrow (0=0 \rightarrow \underline{K})) \neq 1$.

Kripke-style Possible Worlds Semantics

This time we start not with the set **R** of all categorical fixed points meeting a certain condition: R ∈ **R** must be "transparent" in the sense that substituting T[C] for C always preserves semantic value in R. (Note that each opaque fixed point has a least transparent extension, obtained by closing under both directions of <φ(T[C]),v> ∈ V iff <φ(C),v> ∈ V and (1)-(10).) <**R**, ∧, ∨> is a lattice under the same operations as before. Each R ∈ **R** has a dual in **R** that turns gaps into gluts and vice versa, leaving everything else alone. In particular **R** contains a maximal transparent fixed point making each conditional statement both true and false. (Woodruff 1984).

The idea is to develop an increasing series of fixed points P^{β} in tandem with a decreasing series $\mathbf{R}^{\beta}$ of sets of fixed points. A→B is to be true in $P^{\alpha+1}$ iff each $R \in \mathbf{R}^{\alpha}$ validates it. $\mathbf{R}^{\alpha+1}$ is obtained by purging $\mathbf{R}^{\alpha}$ of any valuations that conflict with $P^{\alpha+1}$. P^{α} grows as $\mathbf{R}^{\alpha}$ shrinks, because the smaller $\mathbf{R}^{\alpha}$ is, the easier it becomes for all its members to validate $\underline{A}\rightarrow\underline{B}$. $\mathbf{R}^{\alpha}$ shrinks as P^{α} grows, because the more opinionated a valuation is, the more it comes into conflict with other valuations.

Above we said that $R(\underline{A}) \leq R(\underline{B})$ just in case R assigns 1 to $\underline{B}$ if to $\underline{A}$, and 0 to $\underline{A}$ if to $\underline{B}$. And we said that R validates $\underline{A}\rightarrow\underline{B}$ iff $R(\underline{A}) \leq R(\underline{B})$. Now we further stipulate that $R(\underline{A}) > R(\underline{B})$ iff $R(\underline{A}) \geq R(\underline{B})$ and $R(\underline{A}) \neq R(\underline{B})$. Explicitly,

> $R(\underline{A}) > R(\underline{B}) =_{df}$
>
> R assigns 1 to $\underline{A}$ if to $\underline{B}$, and 0 to $\underline{B}$ if to $\underline{A}$, and (either) 1 to $\underline{A}$ only or 0 to $\underline{B}$ only.

When $R(\underline{A}) > R(\underline{B})$, we say that R *invalidates* $\underline{A}\rightarrow\underline{B}$. The induction goes as follows. Suppose we have P^{α} and $\mathbf{R}^{\alpha}$ in hand.

$$P^{\alpha+1} =$$

$$(\{<\underline{A}\rightarrow\underline{B},1> \mid \text{each } R \in \mathbf{R}^{\alpha} \text{ validates } \underline{A}\rightarrow\underline{B}\}$$

$$\{<\underline{A}\rightarrow\underline{B},0> \mid \text{each } R \in \mathbf{R}^{\alpha} \text{ invalidates } \underline{A}\rightarrow\underline{B}\})^{*}$$

Valuations are <u>incompatible</u> when they disagree in their unique assignments, i.e., one makes a sentence uniquely true (false) that the other fails to make uniquely true (false)..

$$\mathbf{R}^{\alpha+1} =$$

$$\{R \in \mathbf{R}^{\alpha} \mid R \text{ is compatible with } P^{\alpha+1}\}.$$

Limit stages are similar. Suppose we have P^{γ} and $\mathbf{R}^{\gamma}$ in hand for all $\gamma < \lambda$.

$$P^{\lambda} =$$

$$(\{<\underline{A}\rightarrow\underline{B},1> \mid \exists\beta<\lambda\ \forall\gamma\in[\beta,\lambda)\text{ each }R\in\mathbf{R}^{\gamma}\text{ validates }\underline{A}\rightarrow\underline{B})\}\ \cup$$

$$\{<\underline{A}\rightarrow\underline{B},0> \mid \exists\beta<\lambda\ \forall\gamma\in[\beta,\lambda)\text{ each }R\in\mathbf{R}^{\gamma}\text{ invalidates }\underline{A}\rightarrow\underline{B}\ \})^{*}$$

$$\mathbf{R}^{\lambda} = \{R \in \cap_{\gamma<\lambda}\mathbf{R}^{\gamma} \mid R\text{ is compatible with }P^{\lambda}\}.$$

This gives us the sequence depicted in Figure 3:

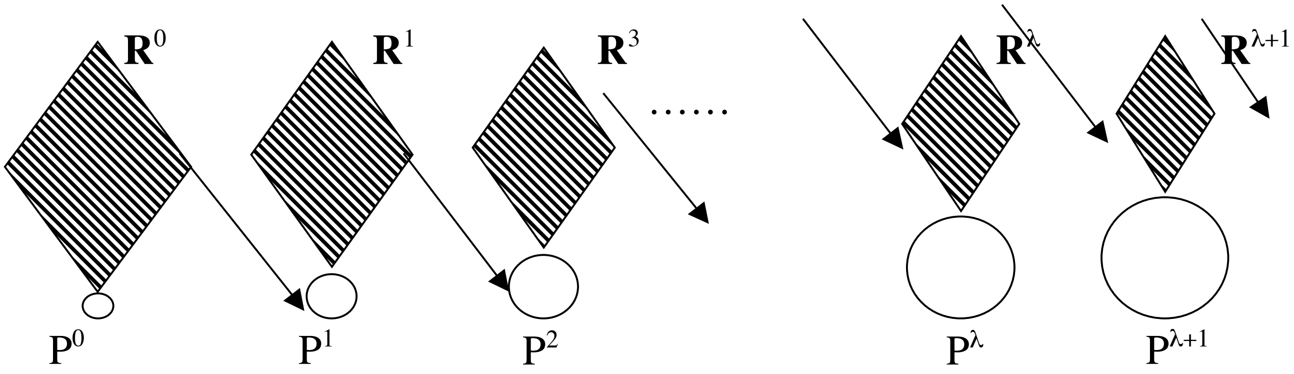


Figure 3

That the $\mathbf{R}^{\alpha}$s are (weakly) decreasing is clear from their definition.

Since P^α varies inversely with $\mathbf{R}^\alpha$, $P^\alpha \subseteq P^\beta$ if $\alpha < \beta$. By the usual cardinality argument, we eventually reach $P^\kappa = P^{\kappa+1}$ and $\mathbf{R}^\kappa = \mathbf{R}^{\kappa+1}$; these are henceforth known as P^∞ and $\mathbf{R}^\infty$. Ultimate values are the values assigned by P^∞. ($||\ldots||$ and P^∞ are one and the same set of ordered pairs.)

****

Some properties of P^∞, aka $||\ldots||$, are clear from the construction. All the P^αs, P^∞ included, are fixed points of Kripke's jump operator, which is just to say they satisfy (1)-(10). What sets P^∞ apart is that it is a fixed point of the operator taking P^α to $P^{\alpha+1}$. If we think of the members of $\mathbf{R}^\infty$ as the worlds accessible from P^∞, the truth-conditions for conditionals are:

(11")

$P^\infty(\underline{A} \rightarrow \underline{B}) = 1$ iff

$\underline{A}\rightarrow\underline{B}$ is validated by all accessible worlds.

(12")

$P^{\infty}(\underline{A} \rightarrow \underline{B}) = 0$ iff

$\underline{A}\rightarrow\underline{B}$ is invalidated by all accessible worlds

Given the definition of an accessible world, this means that $\|\underline{A}\rightarrow\underline{B}\|$ is 1 or 0 according to whether $\underline{A}\rightarrow\underline{B}$ is validated or invalidated by all transparent fixed points compatible with $\|\ldots\|$.[20]

[20] Someone might worry that P^{∞} is inconsistent, that is, assigns some sentence both 1 and 0. It is certainly not consistent if $\mathbf{R}^{\infty}$ is empty, for then every conditional is validated and invalidated by every possible world. If $\mathbf{R}^{\infty}$ has even one member, though, then conditionals have at most one truth-value, hence P^{∞} is consistent overall. And each $\mathbf{R}^{\alpha}$ does have at least one member, since it contains at a minimum P^{α}.

A third property clear from the construction is that P^{∞} verifies every T-biconditional. The proof that $\|T[\underline{C}] \leftrightarrow \underline{C}\| = 1$ is that both directions are validated by all accessible worlds, given that worlds are fixed points. Interchangeability of $T[\underline{C}]$ and $\underline{C}$ in all contexts is proved by induction on the complexity of $\underline{C}$. The one nontrivial part is to show that $\underline{C}$ and $T[\underline{C}]$ are substitutable in $\underline{A}\rightarrow\underline{B}$ if in $\underline{A}$ and $\underline{B}$. Here we rely on the fact that the valuations in $\mathbf{R}^{\infty}$ are transparent. Suppose, for instance, that $\underline{A}\rightarrow(\underline{C}\rightarrow\underline{D})$ is 1 and consider $\underline{A}\rightarrow(T[\underline{C}]\rightarrow\underline{D})$. $\|\underline{A}\rightarrow(\underline{C}\rightarrow\underline{D})\| = 1$ only if for all accessible worlds R, $R(\underline{A})\leq R(\underline{C}\rightarrow\underline{D})$. Transparency tells us that $R(\underline{C}\rightarrow\underline{D}) = R(T[\underline{C}]\rightarrow\underline{D})$, so for all accessible R, $R(A)\leq R(T[\underline{C}]\rightarrow\underline{D})$, whence $\|\underline{A}\rightarrow(T[\underline{C}]\rightarrow\underline{D})\| = 1$.

Next we consider the issues of arbitrariness, ungroundedness, and “laxity” (insufficient strictness). Our example of arbitrariness was the conditional Truthteller $\underline{F}$ $(= (\underline{A}\rightarrow\underline{A})\rightarrow T[\underline{F}])$. Field gives this a value of 1 even though 0 seems on the face of it just as justified.

What value does P^∞ assign? First let's show that $<\underline{F},1>\notin P^1$. $\mathbf{R}^0$ contains a least R assigning 1 to both $\underline{A}\rightarrow\underline{A}$ and T[$\underline{F}$]; call it R_1. $\mathbf{R}^0$ also contains a least R assigning 1 to $\underline{A}\rightarrow\underline{A}$ and 0 to T[$\underline{F}$]; call it R_2. Since R_1 validates $\underline{F}$ and R_2 invalidates it, P^1 leaves ($\underline{A}\rightarrow\underline{A}$)$\rightarrow$T[$\underline{F}$] unevaluated. Since P^1 assigns 1 to $\underline{A}\rightarrow\underline{A}$ (why?) and nothing to $\underline{F}$, P^1 is compatible with both R_1 and R_2. Hence R_1 and R_2 survive into $\mathbf{R}^1$, which by the same argument as before means that P^2 leaves ($\underline{A}\rightarrow\underline{A}$)$\rightarrow$T[$\underline{F}$] unevaluated. Continuing in this way we see that $P^\infty(\underline{F}) = \{\}$.

Our example of ungroundedness was the infinite Curryesque sequence $\underline{K}_i$ = T[$\underline{K}_i$] $\rightarrow$T[$\underline{K}_{i+1}$]. $\mathbf{R}^0$ contains for each $\underline{i}$ a smallest valuation $R_{\underline{i1}}$ assigning 1 to $\underline{K}_{\underline{i}}$ and 1 to $\underline{K}_{\underline{i+1}}$, and a smallest valuation $R_{\underline{i2}}$ assigning 1 to $\underline{K}_{\underline{i}}$ and 0 to $\underline{K}_{\underline{i+1}}$ The first validates $\underline{K}_{\underline{i}}$ and the second invalidates it, so $P^1(\underline{K}_i) = \{\}$ for all $\underline{i}$. But then as above $R_{\underline{i1}}$. and $R_{\underline{i2}}$ survive into $\mathbf{R}^1$, so $P^2(\underline{K}_i) = \{\}$ for all $\underline{i}$, and so on ad infinitum. . The $\underline{K}_{\underline{i}}$s are 1 on the Field semantics but are neither

1 nor 0 by the lights of P^{∞}. The strictness issue was addressed already by the previous semantics, and the same considerations apply. Consider for instance $T'\underline{K}' \rightarrow \neg T'\neg\underline{K}'$. This has ultimate value 1 because every fixed point (including those making $\underline{K}$ both 1 and 0) validates it. $T'\underline{K}' \rightarrow T'\neg\underline{K}'$ is not 1 because fixed points assigning just 1 to $\underline{K}$ fail to validate it, and not 0 either because fixed points assigning just 0 to $\underline{K}$ fail to invalidate it.

****

Look back now at the Top Ten Excellent Features of Field's system (the original system, without possible worlds). How many of these are shared by the system sketched in the last section? I have already said that the Kripke system verifies all T-biconditionals (that's 1.), and permits the substitution salva valutate of $T[\underline{A}]$ for $\underline{A}$ (that's 2.). Certainly it provides an explicit model (that's 3.) which means no hidden paradoxes (that's 4.).[21] When it comes to 7. and 8., there is a tradeoff. The Kripke system gives → a more natural semantics -- a fixed point semantics. -- but at the cost of a less horseshoe-like logic. For a sense of what is lost (and what is not), some key axioms, rules, and metarules of the Field system are

A1 $\models \underline{A}\rightarrow\underline{A}$,

A2 $\models \neg\neg\underline{A}\rightarrow\underline{A}$,

A3 $\models \underline{A}\rightarrow(\underline{A}\vee\underline{B})$,

A4 $\models \underline{A}\wedge\underline{B} \rightarrow \underline{A}$

A5 $\models (\underline{A}\rightarrow\neg\underline{B})\rightarrow(\underline{B}\rightarrow\neg\underline{A})$

A6 $\models (\underline{A}\rightarrow\neg\underline{A})\rightarrow\neg(\top\rightarrow\underline{A})$

R1 $\underline{A}, \underline{A}\rightarrow\underline{B} \models \underline{B}$

R2 $\underline{A}, \neg\underline{B} \models \neg(\underline{A}\rightarrow\underline{B})$

[21] One might worry about inconsistencies, but see the previous note.

R3 $\underline{A} \models \underline{B}\to\underline{A}$

R4 $(\underline{A}\to\underline{B}) \wedge (\underline{A}\to\underline{C}) \models \underline{A} \to (\underline{B} \wedge \underline{C})$

R5 $(\underline{A}\to\underline{C}) \wedge (\underline{B} \to\underline{C}) \models (\underline{A} \vee \underline{B}) \to \underline{C}$

R6 $\underline{A}\to\underline{B} \models (\underline{C}\to\underline{A})\to(\underline{C}\to\underline{B})$.

A1 – A4 are valid on the present semantics, but A5 and A6 hold in rule form only: $\underline{A}\to\neg\underline{B} \models \underline{B}\to\neg\underline{A}$ and $\underline{A}\to\neg\underline{A} \models \neg(T\to\underline{A})$. R1-R5 are truth-preserving but R6 fails.[22] This last is important to the proof that, as in the Field system, → "becomes" horseshoe in bivalent contexts (that's feature 9).[23]

[22] One key weakness of the present logic is its obliviousness in many cases to embedded conditionals. (As evidenced, for instance, in the failure A5 and A6.) It might help to impose more conditions on $\mathbf{R}^0$ than just transparency.

[23] Field's "Theorem on → and ⊃" (Field 2003a) still holds when $\vdash_{\text{LCC}}$ is replaced by $\models$, understood as 1-preservingness in the Kripke semantics. The modified theorem states that:

(ia) $\underline{A}\supset\underline{B} \models \underline{A}\to\underline{B}$

(ib) $(\underline{A}\vee\neg\underline{A})\wedge(\underline{B}\vee\neg\underline{B}) \models (\underline{A}\supset\underline{B})\to(\underline{A}\to\underline{B})$

(iia) $(\underline{A}\vee\neg\underline{A})\wedge(\underline{A}\to\underline{B}) \models \underline{A} \supset \underline{B}$

(iib) $(\underline{A}\vee\neg\underline{A})\wedge(\underline{B}\vee\neg\underline{B}) \models (\underline{A} \to \underline{B})\to(\underline{A}\supset\underline{B})$

Clearly $\underline{B} \models \underline{A} \to \underline{B}$, and $\neg\underline{A} \models \underline{A}\to\underline{B}$. $\vee$-elimination holds because $\|\underline{A} \vee \underline{B}\| = 1$ only if $\|\underline{A}\| = 1$ or $\|\underline{B}\| = 1$. By $\vee$-elimination, $\neg\underline{A}\vee\underline{B} \models \underline{A}\to\underline{B}$, which is the same as (ia). For (ib) and (iib), if $\|(\underline{A}\vee\neg\underline{A})\wedge(\underline{B}\vee\neg\underline{B})\| = 1$ then $\|\underline{A}\| = 1$ or 0, and $\|\underline{B}\| = 1$ or 0. But then $R(\underline{A}) = \|\underline{A}\|$ and $R(\underline{B}) = \|\underline{B}\|$ for all $R\in\mathbf{R}^\infty$. It follows that $\|\underline{A}\to\underline{B}\| = \|\underline{A} \supset \underline{B}\| = 1$ or 0. Either way $R(\underline{A} \supset \underline{B}) = R(\underline{A}\to\underline{B})$ for all $R\in\mathbf{R}^\infty$, so $\|(\underline{A}\supset\underline{B})\to(\underline{A}\to\underline{B})\| =1$ (proving (ib) and $\|(\underline{A}\to\underline{B})\to(\underline{A}\supset \underline{B})\| = 1$ (proving iib). For (iia), that $\|\underline{A}\vee\neg\underline{A}\| = 1$ means that $\|\underline{A}\|$ is 1 or 0. If the latter then $\|\underline{A} \supset \underline{B}\| = 1$. If the former then $\forall R\in\mathbf{R}^\infty\ R(A) = 1$. Also though $\forall\ R\in\mathbf{R}^\infty\ R(\underline{A})\leq R(\underline{B})$ since $\|\underline{A}\to\underline{B}\| = 1$. So $\|\underline{B}\| = 1$, hence $\|\underline{A} \supset \underline{B}\| = 1$. QED

That leaves 5. and 6.: high degree of revenge-immunity, and not because vengeance-threatening notions are inexpressible. One of the great attractions of Field's theory is the astonishing mileage he gets out of a language-internal determinacy operator $D\underline{A}$, defined as $(0=0\rightarrow\underline{A})$ & $\underline{A}$. $D\underline{A}$ is 1 at stage $\alpha+1$ iff $\underline{A}$ is true both at stage $\alpha+1$ and stage α, and 0 at stage $\alpha+1$ iff $\underline{A}$ fails either at the given stage or the one before it. $D\underline{A}$ is 1 at limit stage λ iff $\underline{A}$ is 1 through the closed interval $[\beta, \lambda]$ for some $\beta < \lambda$, and 0 at stage λ if $\underline{A}$ is either 0 at λ or less than 1 through the half-open interval $[\beta, \lambda)$. If $\underline{L}$ is the Liar then $D\underline{L}$ is 0; if $\underline{L}_1$ is the strengthened Liar "I am not determinately true" then $DD\underline{L}_1$ is 0; if $\underline{L}_2$ is the double-strength Liar "I am not determinately determinately true" then $D^3\underline{A}$ is 0; and so on until the ordinal notations run out. That none of these Liars is unevaluable, and indeed each is truly describable as defective$_\alpha$ $(\neg D^\alpha\underline{A} \wedge \neg D^\alpha\neg\underline{A})$ for suitably large α, might seem to support a claim of revenge-immunity.

Distinguish two questions, however. Question #1: is the language able to characterize as defective every sentence that deserves to be so characterized? Question #2: are there intelligible semantic notions such that paradox is avoided only because those notions are not expressible in the language? Field goes a long way towards addressing the first of these, but revenge-mongers have traditionally been more interested in second.

Now it may seem that Field does address the second question, in the section called "Revenge (2)." The revenge-monger (RM) asks us to imagine the chaos that would result if having an ultimate value other than 1 were expressible in the language. A Superliar $\underline{L}_\infty$ could then be constructed saying that $||\underline{L}_\infty|| \neq 1$.. "Why imagine it?," Field asks. The mentioned predicate

> needn't even be added to the language: [it is] already there, at least if the base language from which we started the

> construction…. included the language of set theory, for the construction…showed how to explicitly define [it] in set-theoretic terms … (Field 2003b,??31-2??)

That Field goes so far as to lend RM his tools makes the next step all the more crushing:

> Can't we then reinstitute a paradox? No we can't…our construction yields the consistency of the claim True ($\|\underline{A}\| \neq 1$) $\leftrightarrow$ $\|\underline{A}\| \neq 1$ (ibid.??32??)

The problem is that by Tarski's theorem, “we can't within classical set theory define any notion of determinate truth that fully corresponds to the intuitive notion” (ibid,??33??), what Tarski would have called the metalinguistic notion. The paradox fails because the predicate '$\|\ldots\| \neq 1$' does not mean what RM wanted it to.

But of course it was clear from the start that no truth-predicate definable in the language was going to express the notion of having a semantic value $\neq$ 1. What then was the point of inviting RM into his workshop and offering the use of his tools? I will not speculate about motive, but the result was to get RM off the street where he might really have caused some trouble. From that external vantage point, RM would have seen the old familiar bargain at work and raised the old familiar alarm: consistency is being maintained through a sacrifice of expressive power. I am not saying I agree with RM about this, just that he should be allowed the platform from which his type has traditionally threatened revenge.

Field does have a response to this: "RM, you are taking the semantics and its classical setting too seriously. The most a classical semantics can accomplish is to provide a consistency proof for the theory. The "real" semantics, if there were going to be one, would be carried out in a non-classical set theory, a theory that is needed anyway to deal with the set paradoxes. Sentences like '||this very sentence|| ≠ 1' will not bother us any longer, when we are free of the requirement that $||\underline{A}||=1$ v $||\underline{A}||\neq 1$." A unified treatment of the set and semantic paradoxes is the grail cup of antinomy studies. Russell thought he had it in the theory of types, but we know how that turrned out. How Field's attempt at a unified treatment will turn out, we don't know. But the attempt bears watching.[24]

Bibliography

[24] See Field 2003b.

Barwise, J. & Etchemendy, J. 1987: The Liar: An Essay in Truth and Circularity. Oxford: Oxford University Press

Brady, R. 1989. "The non-triviality of dialectical set theory," in Priest et al 1989, 437-470

Burge, T. 1979: “Semantic Paradox,” Journal of Philosophy 76, 169-98; reprinted in Martin 1984, 83-117

Feferman, S. 1984. "Towards Useful Type-Free Theories, I" Journal of Symbolic Logic 49, 75-111; reprinted in Martin 1984, 237-287

Field, H. 2002. “Saving the Truth Schema from Paradox", *Journal of Philosophical Logic* 31, 1-27

Field, H. 2003a. “A Revenge-Immune Solution to the Semantic Paradoxes” Journal of Philosophical Logic

Field, H. 2003b. "The Semantic Paradoxes and the Paradoxes of Vagueness", this volume

Field, H. forthcoming. "The Consistency of the Naïve (?) Theory of Properties"

Gupta, A. 1982: “Truth and Paradox,” Journal of Philosophical Logic 11, 1-60; reprinted in Martin 1984, 175-235

Herzberger, H. 1982: “Notes on Naive Semantics,” Journal of Philosophical Logic 11, 61-102; reprinted in Martin 1984, 133-174

Kripke, S. 1975: “Outline of a Theory of Truth,” Journal of Philosophy 72, 690-716; reprinted in Martin 1984, 54-81

Martin, R. & Woodruff, P. 1975: “On Representing 'True-in-L' in L,” Philosophia 5, 213-17; reprinted in Martin 1984, 47-51

Martin, R. 1984: Recent Essays on Truth and the Liar Paradox. Oxford: Clarendon

Priest, G. 1979: “The Logic of Paradox,” Journal of Philosophical Logic 8, 219-41

Priest, G., R. Routley, and J. Norman. 1989. Paraconsistent Logic: Essays on the Inconsistent (Philosophia Verlag)

Skyrms, B. 1984: “Intentional Aspects of Semantical Self-Reference,” in Martin 1984, 120-131

Tarski, A. 1983: “The Concept of Truth in Formalized Languages,” in J.H. Woodger, ed. Logic, Semantics, Metamathematics. Indianapolis: Hackett, 152-278

Woodruff, P. 1984: “Paradox, Truth, and Logic (I),” Journal of Philosophical Logic 13, 213-52

Yablo, S. 1985: “Truth and Reflection,” Journal of Philosophical Logic 14, 297-349

Yablo, S. 1993: "Hop, Skip and Jump: The Agonistic Conception of Truth," Philosophical Perspectives 7: 371-396.