THE REVIEW OF SYMBOLIC LOGIC
Volume 13, Number 4, December 2020

INEFFABILITY AND REVENGE

CHRIS SCAMBLER
Department of Philosophy, New York University

Abstract. In recent work Philip Welch has proven the existence of 'ineffable liars' for Hartry Field's theory of truth. These are offered as liar-like sentences that escape classification in Field's transfinite hierarchy of determinateness operators. In this article I present a slightly more general characterization of the ineffability phenomenon, and discuss its philosophical significance. I show the ineffable sentences to be less 'liar-like' than they appear in Welch's presentation. I also point to some open technical problems whose resolution would greatly clarify the philosophical issues raised by the ineffability phenomenon.

§1. Introduction. In [2] and [4], Hartry Field sets out a paracomplete solution to the paradoxes of truth that he claims is *revenge immune*.

Part of the basis for this claim is his theory's provision of a transfinite hierarchy of *definable determinacy operators* $D^{\dot{\alpha}}$.[1] DA is intended to mean that A is determinately true; DDA (or D^2A) that A is determinately determinately true; and so on. These D^αs are associated with theorems to the effect that certain 'intuitively paradoxical' sentences are α-indeterminate for some α. The standard liar Q_0, for example, is equivalent to its own untruth ($\neg T^\ulcorner Q_0 {}^\urcorner$), which implies (in Field's logic) that it isn't determinate ($\neg DQ_0$) by elementary reasoning. In fact one can argue in general that where Q_α is equivalent to its own α-indeterminacy, expressed by $\neg D^\alpha T^\ulcorner Q_\alpha {}^\urcorner$, we always have $\neg D^{\alpha+1} Q_\alpha$, that is, that Q_α is $\alpha + 1$-indeterminate.

The Q_αs are a natural class of paradoxical sentences, and the arguments for their $\alpha + 1$–indeterminacy are simple. One can create more and more complex paradoxical sentences working in Field's object language by e.g., iterated combinations of variants on Curry and liar sentences, using the D operator. It is natural to wonder whether *every* 'intuitively paradoxical' sentence in Field's language can be shown to be α-indeterminate (for some α) in this way, though it should be said that even formulating this idea precisely is not a straightforward matter. (There will be some discussion on this below.)

In recent work Philip Welch has turned his attention to these issues and derived some interesting and philosophically noteworthy results. In particular he has shown that, over any given (standard) ground model M, there will be sentences W in the language of Field's theory whose "ultimate value" is $\frac{1}{2}$ in the Field expansion of M,[2] but for which every definable determinacy operator D^α over M has the ultimate value of $D^\alpha W$ $\frac{1}{2}$ as well. As

Received: September 30, 2018.
2010 *Mathematics Subject Classification*: 03A99.
Key words and phrases: paradoxes, revenge paradoxes, non-classical logic, saving truth from paradox.

[1] I use $\dot{\alpha}$ in the paper to denote a notation for the ordinal α in a given formal language. Generally I may leave of the dots when notation systems aren't explicitly relevant.

[2] For 'Field expansion', I mean a Strong Kleene expansion of M for a language with conditional operator and truth predicate as described in [4]. §2 has more details.

© Association for Symbolic Logic, 2018
doi:10.1017/S1755020318000473

https://doi.org/10.1017/S1755020318000473 Published online by Cambridge University Press
CrossMark

a result none of these sentences or their negations get a designated value in the Field expansion of M. Welch has dubbed these sentences *ineffable liars*, suggesting that they are examples of liar-like sentences about which the object language of Field's theory must remain silent. Welch goes on to suggest that the existence of such ineffable sentences may bear on the revenge immunity of Field's logic.

In this paper I present a slightly more general characterization of ineffability to the one given in [9], and a simpler and slightly corrected argument for the existence of ineffable liars.[3] Along the way, we will see that the sentences are less 'liar-like' than they may initially seem. I then discuss the significance of these results with respect to the revenge immunity of Field's theory, pointing toward a couple of open logical problems that seem to be of special philosophical significance in this regard.

First I will give some background.

§2. Field's construction. In [2] and [4], Field shows how to expand any (standard) model M for a classical language $\mathcal{L}$ to a model M^+ for the language $\mathcal{L} \cup \{\rightarrow, T\}$, which I will hereafter call $\mathcal{L}^+$, satisfying naive principles of truth and validating reasonable laws for $\rightarrow$ (with salient exceptions: conditional proof, and contraction). He then defines a validity relation $\models_F$ for the expanded language using the expanded models. I will assume familiarity with Field's construction and logic, but must say a few things for later reference.

All formulas of the form $A \rightarrow B$ are atomic in $\mathcal{L}^+$. The construction proceeds by iterating rounds of the Kripke construction[4] (generating strong Kleene valuations for T, holding values for $\rightarrow$ fixed) and successively refining the valuations for $\rightarrow$ in the light of information coming from the iterated rounds of the Kripke construction. If F is a valuation for conditionals, let $|A|_F$ be the value for A in the least Kripke fixed point model expanding M and the valuation F for $\rightarrow$. A sequence F_α of valuations for conditionals can then be inductively defined using:

$$F_\alpha^+ = \{\langle \ulcorner A \rightarrow B \urcorner, 1\rangle : \exists\beta\forall\gamma \in [\beta, \alpha)[|A|_{F_\gamma} \leq |B|_{F_\gamma}]\} \tag{1}$$

$$F_\alpha^- = \{\langle \ulcorner A \rightarrow B \urcorner, 0\rangle : \exists\beta\forall\gamma \in [\beta, \alpha)[|A|_{F_\gamma} > |B|_{F_\gamma}]\}. \tag{2}$$

F_α then consists of F_α^+, F_α^-, and all other conditionals paired with $\frac{1}{2}$. I will henceforth write $|A|_\alpha$ for $|A|_{F_\alpha}$.[5]

There are three broadstrokes possibilities for the sequence $|A|_\alpha$ as α runs through the ordinals: $||A|| = 1$ means $|A|_\alpha$ is eventually constant on 1, m.m. for $||A|| = 0$; $||A|| = \frac{1}{2}$ iff $||A|| \notin \{0, 1\}$. In [2], Field proved

THEOREM 2.1. *There are ordinals* Δ *for which* $|A|_\Delta = ||A||$.

[3] This paper arose because I noticed a small error in Welch's original presentation, discussed in note 19, and Welch and I both set about looking for ways to fix it. Unsurprisingly, my solution diverges more from Welch's original argument than his does. The problem with Welch's original argument, and the relation between his solution to it and mine, will be briefly discussed below, in note 19.

[4] In the sense of [5].

[5] I understand from conversation with Field that he now prefers the definition for the 0-clause for conditionals according to which a conditional gets the value 0 at a stage iff its antecedent had the value 1 and the consequent had the value 0 at the previous stage. All the arguments I make below are unaffected by which approach we take here; the modifications that would be needed to accommodate it are obvious enough that I won't point them out.

https://doi.org/10.1017/S1755020318000473 Published online by Cambridge University Press

Such Δ are called *acceptable points*. These are (as Field shows) unbounded in the ordinals, and I will enumerate them by Δ_α. The least such ordinal Δ_0 varies with the cardinality of the ground model M, and we will have more to say about its precise location in the ordinals relative to the cardinality of M below.

As already mentioned, one of the important facts about Field's construction is that it furnishes a transfinite hierarchy of determinateness operators $D^{\dot\alpha}$, defined whenever a notation $\dot\alpha$ is available for an ordinal α in $\mathcal{L}^+$ (relative to a given ground model).[6] I'll now present some basic definitions and facts about this hierarchy.

The most basic definition is the following:

$$DA := A \wedge \neg(A \rightarrow \neg A). \tag{3}$$

Given the other rules of Field's logic, we have that

$$|\neg(A \rightarrow \neg A)|_\alpha = 1$$

just in case A got the value 1 at the previous stage (where this means on a continuous tail up to α if α is a limit).[7] So, intuitively, $|DA| = 1$ holds at a given stage when A holds at that stage, and held at the previous stage.

Of course, the operator D can be iterated; and how many times it can be iterated depends on the availability of ordinal notations. The following definitions and lemmas carve out a framework of sorts for discussing these.

DEFINITION 2.2. *A two-place formula $P(x, y)$ of $\mathcal{L}^+$* bivalently defines a prewellorder (*over a given ground model M*) *just in case*

- $P^{\leq} = \{\langle a, b\rangle : ||P(a, b)|| = 1\}$ *is a prewellorder*;[8]
- *For all $b \in Fld(P^{\leq})$ and all $a \in M$, $||P(a, b) \vee \neg P(a, b)|| = 1$.*

When P bivalently defines a prewellorder, $P^{\leq}$ will have some definite ordinal type, and consequently one can view the (names of)[9] objects of $Fld(P^{\leq})$ as notations for ordinals in $\mathcal{L}^+$ (relative to the ground model M). If $a \in Fld(P^{\leq})$ has a set of P-predecessors of order type α, I will say that a is a *notation* for α and write $\dot\alpha$ for a (leaving P, M clear from context).[10]

Suppose we are given a P that bivalently defines a prewellorder. By recursion along P, and working in $\mathcal{L}^+$, we can define an operator $D_P(x, y)$ on (codes of) formulas and notations by:

---

[6] By 'relative to a given ground model', I will always mean 'in the Field construction over a given ground model'; in this case, specifically what is meant it that $\dot\alpha$ is a notation for α in the Field expansion of whatever ground model we are considering. To avoid such a mouthful, I will generally use the shorter phrase 'relative to a given ground model'.

[7] Because $A \rightarrow \neg A$ gets value 0 at α just in case there is a β such that for all $\gamma \in [\beta, \alpha)$ $|A|_\gamma$ is greater than $1 - |A|_\gamma$ (by the strong Kleene clause for negation), which happens just when $|A|_\gamma$ is 1; and when α is a successor, this reduces to A getting the value 1 at the previous stage.

[8] Here I assume that a and b name themselves in the language for all a, b in the domain M. This means that technically the language varies with the ground model. Of course this could be avoided, but the convention adopted seems to me the simplest.

[9] See previous note.

[10] One annoying technical detail here is that a notation for an ordinal may only 'stabilize' as such fairly late in the construction. A full definition of notation would proceed level by level, and would have the definition above as a special case. For present purposes, I have chosen to give this looser version of the definition and simply restrict attention to sufficiently late stages of the construction, because the alternative seems needlessly complicated.

https://doi.org/10.1017/S1755020318000473 Published online by Cambridge University Press

$$D_P(A, \dot{0}) = A \tag{4}$$

$$D_P(A, \alpha \dot{+} 1) = D(D_P(A, \dot{\alpha})) \tag{5}$$

$$D_P(A, \dot{\lambda}) = \forall x[P(x, \dot{\lambda}) \rightarrow TD_P(A, x)]. \tag{6}$$

The notions of $\dot{0}$ and $\cdot \dot{+} 1$ can obviously be written out in full in terms of P in $\mathcal{L}^+$. Really, D_P operates directly on the *codes* of formulas, not formulas; this is an object language definition. I have left out the corner quotes for readability.[11] The formula (coded by) $D_P(A, \dot{\alpha})$ is what I have been referring to as $D^{\dot{\alpha}}A$, and intuitively is the αth iterate of the D operator applied to A. Importantly, one can show that if the order type of P is less than or equal to that of Q, then the operators defined by P 'agree' with those of the corresponding type as defined by Q, meaning that we can in effect 'forget about' which bivalently defined prewellorder we employ. Similarly, since P is a prewellorder, there may and in general will be different notations relative to P for each ordinal $< OT(P^{\leq})$, but again we can prove it is a matter of indifference which is used. See [3] for further discussion of these operators.

Recall that DA is true at a given stage when A is true at that stage, and was true at the last stage. One might therefore expect DDA to 'look two stages back', and so on for further iterations. This is the content of the following lemma.

LEMMA 2.3. *Let $\dot{\alpha}$ be notation for α over some ground model M. Then, for any A and for sufficiently large γ (see note 10), $|D^{\dot{\alpha}}A|_\gamma = 1$ if and only if there is a σ with $\sigma + \alpha \leq \gamma$, and $|A|_\beta = 1$ for every $\beta \in [\sigma, \gamma]$.*

Proof. A routine induction on α. □

In producing ineffable liars, it will also be important to have conditions relating $||D^{\dot{\alpha}}A||$ to $|A|_\gamma$ as γ goes through the ordinals. The case of 1 is straightforward: we have

LEMMA 2.4. $||D^{\dot{\alpha}}A|| = 1$ *iff* $||A|| = 1$.

Proof. An easy consequence of Lemma 2.3. □

But it is the case of 0 and $\frac{1}{2}$ that we are really interested in; in this case, it is possible for $||A|| = \frac{1}{2}$, but $||D^{\dot{\alpha}}A|| = 0$, if certain conditions are met. The following series of lemmas is designed to clarify these conditions. In so doing we will arrive at necessary and sufficient conditions for a sentence to be ineffable.

LEMMA 2.5 (Diagonal Lemma). *If $|A|_\gamma \neq 1$, then for each $\alpha > 0$, $|D^{\alpha \dot{+} 1}A|_{\gamma + \alpha} = 0$ so long as the relevant notation exists.*

Proof. By induction on α.[12] For the base case if $|A|_\gamma \neq 1$ it is easy to verify that $|DA|_{\gamma+1} = 0$, whence $|D^2A|_{\gamma+1} = 0$ as the lemma requires. (Indeed, for all finite n we will have the stronger $|D^nA|_{\gamma+n} = 0$.)

[11] To put the corner quotes back in, you need to write $\ulcorner A \urcorner$ wherever A is written in the definition, and also put corner quotes around the entire expressions on the right hand side of the identities (5), (6). For the purposes of the definition, the bare operator D must also be understood as operating on the code of A to yield a code for $A \wedge \neg(A \rightarrow \neg A)$. The actual formulas can of course be decoded from that version of the definition in an effective way given any reasonable coding function $\ulcorner \cdot \urcorner$.

[12] Because we are simply assuming notations exist for all relevant α, I'll drop the dots in this argument. Generally I may leave them off where no confusion is possible. I will also repeatedly make use of the fact (which should be evident from the definitions) that $|D^\alpha A|_\gamma = 0$ implies $|D^\beta A|_\gamma = 0$, all $\beta > \alpha$.

https://doi.org/10.1017/S1755020318000473 Published online by Cambridge University Press

Suppose $\alpha = \beta+1$, and that it holds for β. Then we have $|D^{\alpha}A|_{\gamma+\beta} = 0$ by the inductive hypothesis, so that

$$|\neg(D^{\alpha}A \rightarrow \neg D^{\alpha}A)|_{\gamma+\alpha} = 0,$$

which implies $|D^{\alpha+1}A|_{\gamma+\alpha} = 0$, as desired.

Suppose $\alpha = \lambda$, and that it holds for all $\beta < \lambda$. Then we have $|D^{\beta+1}A|_{\gamma+\beta} = 0$ for each $\beta < \lambda$ by the inductive hypothesis, and this implies that $|D^{\lambda}A|_{\gamma+\beta} = 0$ for all such β. This in turn implies that $|\neg(D^{\lambda}A \rightarrow \neg D^{\lambda}A)|_{\gamma+\lambda} = 0$, so that $|D^{\lambda+1}A|_{\gamma+\lambda} = 0$, as required. □

REMARK 2.6. In the context of the last lines of the previous proof, note that we do not necessarily have $|D^{\lambda}A|_{\gamma+\lambda} = 0$; indeed, if $|A|_{\delta} = 1$ for all $\delta \in (\gamma, \gamma+\lambda]$ we will have $|D^{\lambda}A|_{\gamma+\lambda} = 1$. (See Table 1.)

Table 1. *Visualization of the 'worst case scenario' for the diagonal lemma, explaining why the simpler* $|D^{\alpha}A|_{\gamma+\alpha} = 0$ *idea goes wrong. At* $\gamma+\omega$, $D^{n}A$ *has the value* 1 *for every n, yielding the value* 1 *for* $D^{\omega}A$ (*emphasized with exclamation points*). *Accordingly, the correct diagonal runs one column to the right.*

S	A	DA	D^2A	D^3A	…	$D^{\omega}A$	$D^{\omega+1}A$	$D^{\omega+2}A$	…	$D^{\alpha}A$	…
γ	<1	?	?	?	…	?	?	?	…	?	…
$\gamma+1$	1	0	0	0	…	0	0	0	…	0	…
$\gamma+2$	1	1	0	0	…	0	0	0	…	0	…
$\gamma+3$	1	1	1	0	…	0	0	0	…	0	…
⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮	…
$\gamma+\omega$	1	1	1	1	…	1!!	0	0	…	0	…
$\gamma+\omega+1$	1	1	1	1	…	1	1	0	…	0	…
$\gamma+\omega+2$	1	1	1	1	…	1	1	1	…	0	…
⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮	…
$\gamma+\alpha-1$	1	1	1	1	…	1	1	1	…	0	…
$\gamma+\alpha$	1	1	1	1	…	1	1	1	…	1	⋱

LEMMA 2.7. $||D^{\dot{\alpha}}A|| = 0$ *iff either*

- $||A|| = 0$

or

- *There is an* ϵ *such that all* $\delta > \epsilon$ *either have* $|A|_{\beta} \neq 1$ *for some* $\beta \in [\delta, \delta+\alpha)$, *or* $|A|_{\delta+\alpha} = 0$.

Proof. For the right to left: This is obvious if $||A|| = 0$, so assume that the values of A are as described in the second bullet. We show that eventually all γ are such that $|D^{\alpha}A|_{\gamma} = 0$.

Pick $\delta > \epsilon$ (ϵ witnessing the existential claim in the second bullet) such that $|A|_{\delta} < 1$. Let $f(\delta)$ be the least ordinal $>\delta$ such that $|A|_{f(\delta)} < 1$. (Our assumption implies f is total on ordinals.) It suffices to show that $|D^{\alpha}A|_{\beta} = 0$ for each $\beta \in (\delta, f(\delta)]$, since then the same reasoning applies in the interval $[f(\delta), f(f(\delta))]$, &c. This is what we aim to show.

There are two cases. Case one: for some $\gamma \in (\delta, \delta+\alpha)$, $|A|_{\gamma} < 1$. Then $\gamma = \delta+\nu$ for some $\nu < \alpha$, so by the diagonal Lemma we must have $|D^{\alpha}A|_{\beta} = 0$ for all $\beta \in (\delta, \delta+\nu]$. Since by hypothesis $|A|_{\delta+\nu} < 1$, we are done.

https://doi.org/10.1017/S1755020318000473 Published online by Cambridge University Press

Case two: $|A|_\beta = 1$ for all $\beta \in (\delta, \delta + \alpha)$. In this case, the diagonal Lemma still guarantees that $|D^\alpha A|_\beta = 0$ for all such β; but then $|A|_{\delta+\alpha}$ must be 0 (since $\delta > \epsilon$). This implies that $|D^\alpha A|_{\delta+\alpha} = 0$, and obviously $|A|_{\delta+\alpha} < 1$, so again we are done.

For the left to right we proceed by contraposition. Suppose that $||A|| \neq 0$, and

$$\forall\epsilon\exists\delta > \epsilon\forall\gamma \in (\delta, \delta + \alpha)[|A|_\gamma = 1, \text{and } |A|_{\delta+\alpha} > 0]. \tag{7}$$

Fix arbitrary ϵ; we will produce $\nu > \epsilon$ at which $|D^\alpha A|_\nu \neq 0$. Let $\delta > \epsilon$ be as in (7). Using Lemma 2.3, we have $|D^\beta A|_{(\delta+1)+\beta} = 1$ for each $\beta < \alpha$. If α is a limit ordinal we get $|D^\alpha A|_{\delta+\alpha} = 1$ immediately as a consequence. Otherwise we have

$$|\neg(D^{\alpha-1}A \to \neg D^{\alpha-1}A)|_{\delta+\alpha} = 1;$$

and since by hypothesis $|A|_{\delta+\alpha} > 0$, we must have $|D^\alpha A|_{\delta+\alpha} > 0$ as required. So take $\nu = \delta + \alpha$. □

LEMMA 2.8. *$||D^{\dot\alpha}A|| = \frac{1}{2}$ iff $||A|| = \frac{1}{2}$, and for all ϵ there is a $\delta > \epsilon$ with $|A|_\beta = 1$ for each $\beta \in [\delta, \delta + \alpha)$, and $|A|_{\delta+\alpha} \neq 0$.*

Proof. A simple consequence of Lemmas 2.4 and 2.7. □

We can now give sharp conditions for a sentence to be ineffable. First, the formal definition:[13]

DEFINITION 2.9. *A sentence W is* ineffable *just in case $||W|| = \frac{1}{2}$ but $||D^{\dot\alpha}W|| = \frac{1}{2}$ for all notations $\dot\alpha$ (relative, of course, to some fixed ground model M).*

Next, the condition for ineffability. Intuitively (but roughly!), what it says is that a sentence is ineffable when it gets ultimate value $\frac{1}{2}$, but takes the value 1 for a sequence of α successive stages arbitrarily late for every α that has a notation.

THEOREM 2.10. *A is ineffable iff $||A|| = \frac{1}{2}$, and for every α with a notation (over M) and all ϵ, there is a $\delta > \epsilon$ with $|A|_\beta = 1$ for each $\beta \in [\delta, \delta + \alpha)$, and $|A|_{\delta+\alpha} \neq 0$.*

Proof. This follows by the definition of ineffability together with Lemma 2.8. □

It follows that in order to prove there is an ineffable sentence, it suffices to (1) show that there are only notations over a given M up to some ordinal ν_M, and (2) then to show that there is a sentence with ultimate value $\frac{1}{2}$ on M which takes the value 1 unboundedly beneath ν_M, and on unboundedly long sequences of stages. We intend to pursue this two part strategy in the next section.

§3. Ineffability. In this section I present a proof that ineffable sentences exist that relies heavily on the methods of Welch in [9], but that is slightly simpler in ways that may turn out to be of some philosophical significance (as I suggest in §4).

In [9], Welch establishes several results relating the Field expansion of M to the M-relative constructible sets. Recall that the latter are defined by:

$$L_0^M = M$$

[13] It is worth noting that Welch himself does not give such a general definition of ineffability: his idea is to find ineffable *liars*, primarily, that are diagonalized sentences. As I hope will become clear (especially in §4), the more general characterization is logically and philosophically useful.

https://doi.org/10.1017/S1755020318000473 Published online by Cambridge University Press

$$L^M_{\alpha+1} = \mathcal{P}_{def} L^M_\alpha$$

$$L^M_\lambda = \bigcup_{\alpha<\lambda} L^M_\alpha.$$

Where $\mathcal{P}_{def}X$ is the set of all subsets of X definable (in the language of set theory) over X with parameters from X. Recall also that M is a Σ_2-*elementary substructure* of N, written $M \prec_{\Sigma_2} N$, just in case any Σ_2 formula ϕ with parameters only from M that is true in N is also true in M.[14]

Probably the most fundamental result of Welch's here, for our purposes at least, is the following.[15]

THEOREM 3.1. *Let M be a structure for $\mathcal{L}$. Then Δ^M_0 (the least acceptable point over M) is the least ordinal ρ_0 such that $L^M_{\rho_0} \prec_{\Sigma_2} L^M_{\rho_1}$ for $\rho_1 > \rho_0$.*

The basic idea of the proof is as follows. Let τ_α enumerate the admissible ordinals (always relative to M, though to keep the notation succinct I will drop reference to M where no confusion is possible; see the note for a brief explanation and further references).[16] Welch begins by proving that one can define the semantic value of any sentence A at stage α over L_{τ_α} by a Σ_2 formula. This is possible because, as is hinted by the syntactic form of (1), (2), the definition of $|A|_\alpha$ is a Σ_2 recursion, and the L-ranks of M-admissible ordinals satisfy enough set theory to carry it out. (Going into full detail would take us too far afield here; a more detailed sketch is in the next note.)[17] For admissible limits of admissibles τ_λ, then, $|A|_\gamma$ is definable by a Δ_0 formula (only bounded quantifiers) over L_{τ_λ} whenever $\gamma < \tau_\lambda$. Since $\rho_0 = \tau_{\rho_0}$ (as Welch shows in [7, Lemma 2.1]), we have that ρ_0 is an admissible limit of admissibles, and hence that $|A|_\gamma$ is Δ_0-definable over $L_{\tau_{\rho_0}} = L_{\rho_0}$ for $\gamma < \rho_0$, and that $|A|_{\rho_0}$ is Σ_2 definable over L_{ρ_0}.

These facts allow us to infer Theorem 3.1. To show that ρ_0 is acceptable it suffices to show that $|A|_{\rho_0} = 1$ if and only if $||A|| = 1$, and similarly for 0. For the right to left direction, note that because $|A|_{\rho_0}$ is Σ_2 definable over L_{ρ_0} and $L_{\rho_0} \prec_{\Sigma_2} L_{\rho_1}$, we must have $|A|_{\rho_0} = |A|_{\rho_1}$ for any sentence A. Hence observe $|A|_{\rho_0+\alpha} = |A|_{\rho_1+\alpha}$ for all α. In particular we find (using $\rho_0 + \rho_1 = \rho_1$) that

$$\forall\alpha[|A|_{\rho_1\cdot\alpha} = |A|_{\rho_0}].$$

---

[14] Recall also that a formula (in the language of set theory) is said to be Δ_0, Π_0 or Σ_0 (indifferently) iff all its quantifiers are bounded; Π_{n+1} (Σ_{n+1}) if it is of the form $\forall x\phi(x, \vec{z})$ ($\exists x\phi(x, \vec{z})$) where $\phi(x, \vec{z})$ is Σ_n (Π_n). Where such complexity classes are used in this paper, I shall always be referring to formulas of the language of set theory, possibly expanded to include the language of whatever ground model is under consideration.

[15] This is Theorem 2.1 of [7].

[16] An ordinal τ is *admissible* over M if L^M_τ is a model of a certain weak set theory (Kripke Platek with Urelements from M). See [1] for more details.

[17] First, because the Kripke construction can be given an inductive definition in the sense of [6], the first level of the construction can be given by a Σ_1 definition over $L^M_{\tau_1}$ (as proven e.g., in Barwise, [1, 215]). But then the next valuation for conditionals can easily be decoded from this, and so if we have a further admissible we can define another round of the Kripke construction with 'access' to that valuation, gnerating the finite stages of the construction (always by a Σ_1 formula over L_{τ_n}, since only the previous valuation for conditionals is relevant). At limits, however, we need to use the Σ_2 clauses corresponding to (1), (2) to get the λth valuation for conditionals, as well as the Σ_1 truth definition, so in general the complexity is Σ_2. More detail can be found in [8] and [9].

https://doi.org/10.1017/S1755020318000473 Published online by Cambridge University Press

Since multiples of ρ_1 are unbounded we must therefore have $||A|| = 1$ implies $|A|_{\rho_0} = 1$.

For the left to right, we need to show that $|A|_{\rho_0} = 1$ implies $||A|| = 1$. In fact we prove this result only for a conditional sentence C; the result then follows for arbitrary formulas by a routine induction, using the fact that the values for arbitrary sentences follow mechanically *via* the Kripke construction. For this in turn it suffices to show (on the assumption $|C|_{\rho_0} = 1$) that $|C|_\gamma = 1$ for all $\gamma \in [\rho_0, \rho_1)$, since the by the reasoning of the previous paragraph the values of formulas will constantly 'loop' in the same pattern from there. Towards the the latter claim, note that $|C|_{\rho_0} = 1$ implies (since C is a conditional) that there is $\delta < \rho_0$ with $|C|_\gamma = 1$ for $\gamma \in [\delta, \rho_0]$; hence it follows that

$$L_{\rho_0} \models \forall \gamma > \delta[|C|_\gamma = 1] \tag{8}$$

using the definability of $|C|_\gamma$ in L_{ρ_0}. Now, suppose for reductio that there is $\gamma \in (\rho_0, \rho_1)$ for which $|C|_\gamma \neq 1$. Then we have

$$L_{\rho_1} \models \exists \gamma > \delta[|C|_\gamma \neq 1],$$

again using the definability result. But this is a Σ_1 assertion true in L_{ρ_1} with parameters from L_{ρ_0}, and so must *reflect down* to be true in L_{ρ_0}, contradicting (8) and completing the reductio. An exactly analogous argument serves to establish the analogous claim for 0, and hence that $|A|_{\rho_0} = ||A||$ for all A. Thus, ρ_0 is an acceptable point. Finally, a similar 'reflection argument' can be given to the effect that ρ_0 is the least such. Thus, we infer that $\rho_0 = \Delta_0$, the least ordinal for which $|A|_\Delta = ||A||$. Similar reasoning serves to establish that $\Delta_\alpha = \rho_\alpha$ for all α, where $\rho_\alpha = \rho_0 + \rho_1 \cdot \alpha$.

This result has two important consequences (both due to Welch again). The first is

COROLLARY 3.2 (Stability mirage). *There is a sentence W such that*:

- $||W|| = \frac{1}{2}$
- *There are unboundedly many pairs of ordinals* $\alpha < \iota < \rho_0$ *such that* (*a*) ι *is additively indecomposable, and* (*b*) $|W|_\beta = 1$ *for every* $\beta \in [\alpha, \iota]$.

Proof. Define $rk(A)$ to be the least ordinal γ (if it exists) such that the semantic value of A is constant on 1 from stage γ on in the Field construction. The notion of $rk(A)$ makes sense when relativized to any stage σ of the (M-)constructible hierarchy: $rk_\sigma(A)$ is the least ordinal $\gamma < \sigma$ (if it exists) at which the semantic value of A is constant on 1 from γ on from the point of view of L_σ (that is, iff $L_\sigma \models rk(A) = \gamma$). $||A|| = 1$, then, if and only if $rk_{\rho_0}(A)$ is defined (and so necessarily is $< \rho_0$).

Corollary 3.2 then follows from the following lemma.

LEMMA 3.3 (Overspill). *There is a sentence W and ordinals*

$$\nu_0 < \nu_1 \in (\rho_0, \rho_1)$$

such that

- $\nu_1 = \tau_{\nu_1}$
- $rk_{\nu_1}(W) = \nu_0$.

The lemma says that W appears to take a constant value on 1 for a sequence of stages of length $\nu_1 > \rho_0$ only starting after ρ_0. Of course this implies that $||W|| = \frac{1}{2}$, since all sentences that actually have absolute value 1 take that value on constantly before ρ_0 and retain it forever more.

https://doi.org/10.1017/S1755020318000473 Published online by Cambridge University Press

For a sketch of the proof: Welch shows that for every $\gamma < \rho_0$ there is a sentence C_γ with $rk(C_\gamma) = \gamma$ [9, Lemma 1.6]. So if the Overspill lemma is false, the following expression

> $x = \tau_x$, and every ordinal α less than x has some sentence S with $rk_x(S) = \alpha$, and moreover for no $y > x$ with $\tau_y = y$ is there a sentence S with $rk_y(S) > x$

would be a Π_1 (and hence Σ_2) definition of ρ_0 over L_{ρ_1} (for *no* $y > x \ldots$), which is impossible given $L_{\rho_0} \prec_{\Sigma_2} L_{\rho_1}$.[18]

But now Corollary 3.2 follows. Taking $\delta < \rho_0 < \nu_0$ we can infer

$$L_{\rho_1} \models \exists \iota[\iota = \tau_\iota \wedge rk_\iota(W) > \delta]$$

since this is a Σ_1 truth over L_{ρ_1} with the parameter δ from L_{ρ_0}, we infer

$$L_{\rho_0} \models \exists \iota[\iota = \tau_\iota \wedge rk_\iota(W) > \delta] \tag{9}$$

and since δ is arbitrary $<\rho_0$, and the ι must be additively indecomposable, the result follows. □

For an intuitive gloss, note that the result implies that for arbitrarily large additively indecomposable ordinals $\iota < \rho_0$, the sentence W appears to have a rank (that is, to take on a value constant on 1 up to ι); but ultimately the sequence of 1s is always interrupted by other values, since $||W|| = \frac{1}{2}$. The sentence W can be construed as a sort of 'stability mirage': as we progress through the construction, it appears to stabilize on value 1 for longer and longer sequences $<\rho_0$, but these always ultimately break down and the sentence ultimately does oscillate unboundedly in its value.

It follows (by Theorem 2.10) that such a sentence will be ineffable so long as we can show that there are only notations for ordinals $<\rho_0$. And, as the reader might have guessed, Welch has proven this, once again by a reflection argument. Indeed, the result is the other "important corollary" to Theorem 3.1 I mentioned:

COROLLARY 3.4. *The longest possible bivalently defined prewellorder over a given ground model has type* $\Delta_0 = \rho_0$ (*relative to M, of course*). *Hence there are only notations* $\dot{\alpha}$ *for* $\alpha < \Delta_0$.

This time I entirely omit the argument. Roughly, the idea is that if there were a bivalently defined prewellordering of type Δ_0, we could show $\Delta_0 = \rho_0$ to be Σ_2 definable over L_{ρ_1} with parameters from L_{ρ_1}, which is a contradiction, as in note 18. See [9] Lemma 1.7 for more details here.

We now have

COROLLARY 3.5. *The sentence W from Corollary 3.2 is ineffable* (*over the ground model M*).

Proof. By Theorem 2.10, Corollary 3.2, and Theorem 3.4. Explicitly: fix an α with a notation over M. By Theorem 3.4, $\alpha < \Delta_0^M$. By Corollary 3.2 there are arbitrarily large $\iota > \nu \in [\alpha, \Delta_0^M)$ such that is ι is additively indecomposable, and for all $\gamma \in [\nu, \iota) |W|_\gamma = 1$. Because ι is additively indecomposable, it follows that there is a proper sub-interval $[\nu, \nu + \alpha]$ on which W takes the value 1. (Hence $|D^\alpha W|_{\nu+\alpha} = 1$.) Since α stands for

[18] Any ordinal L_{ρ_1} definable by a Σ_2 formula without parameters (such as the above) in L_{ρ_1} must belong to L_{ρ_0}, and so in particular is not identical to ρ_0.

https://doi.org/10.1017/S1755020318000473 Published online by Cambridge University Press

an arbitrary notation over M and such ι are unbounded below Δ_0^M, the result follows by Theorem 2.10. □

I will conclude this section with some observations on the nature of ineffability and the relation between the results presented here and those of [9].

When a sentence has ultimate value $\frac{1}{2}$ in the Field construction – so that its value is neither eventually constant on 1 or 0 – its sequence of values will always *loop in a fixed period* α in the following sense: there will be some ordinal α such that $|A|_\delta = |A|_{\delta+\alpha}$ for any sufficiently large δ. For example, the determinacy liars

$$Q_\alpha \equiv \neg D^\alpha T^\ulcorner Q_\alpha {}^\urcorner$$

eventually take the value $\frac{1}{2}$, followed by α 1s, followed by the value $\frac{1}{2}$ again, and so on repeating; such sentences have period $\alpha + 1$ when α is a successor, and α otherwise. As a matter of fact, all of the 'natural' sentences one thinks of with ultimate value $\frac{1}{2}$ can be assigned some such period $< \Delta_0^M$; but W obviously can be assigned no fixed period beneath ρ_0^M. Indeed, since (9) implies

$$L_{\rho_0} \models \forall\delta\exists\iota[\iota = \tau_\iota \wedge rk_\iota(W) > \delta]$$

which is Π_2 over L_{ρ_0}, this must (as Welch observes) reflect up to be true in L_{ρ_1}, i.e.,

$$L_{\rho_1} \models \forall\delta\exists\iota[\iota = \tau_\iota \wedge rk_\iota(W) > \delta]$$

so that W cannot be assigned any period $< \Delta_1^M$. (In fact this will be true for any sentence with no period $< \Delta_0^M$, by essentially this argument.) Necessarily, of course, no sentence can have period $> \Delta_1^M$ in the Field expansion of M, so we conclude that W has period exactly Δ_1^M.

Welch calls such sentences 'sporadic', and uses them in [9] to index special determinateness hierarchies D^C (for sentences C) that can in turn be used to furnish ineffable liars. The indexing by sentences works because the 'stability ordering' on sentences, holding between two sentences when they both ultimately get value 1 (or, as I will say, 'stabilize') but the one takes on its constant value no later than the other, is internally (and bivalently, in the sense of Definition 2.2) definable by a formula in the language of the construction;[19] so any sentence that really stabilizes acts as a notation for an ordinal, and can be used as an index for a determinacy hierarchy. The sentence W above then is then an apparent notation arbitrarily late in the construction, but ultimately is no notation at all; Welch exploits this to show that a diagonalized liar sentence Q_W equivalent to $\neg D^W Q_W$ is ineffable.

[19] It was at this point that the small mistake arose in Welch's original presentation; roughly, Welch originally claimed that there was a formula of the language $P(x, y)$ that took ultimate value 1 if x and y coded sentences that stabilize and x stabilized no later than y, value 0 if they code sentences that stabilize and x stabilized later than y, and $\frac{1}{2}$ otherwise. In fact this is inconsistent. Welch's solution (cf. [10]) is to only have P take the value one on x and y if either y does not stabilize, or x stabilizes before y; P takes the value $\frac{1}{2}$ on x and y in all other cases.

In the process of thinking through the same problem, I arrived at this alternative solution, which does not seem to require the internal definability of the stability hierarchy (or this P). In this respect the argument is simpler. It also has shows that the ineffable sentences can be characterized in terms of their having a distinctive type of semantic value in Field's 'fine-grained' algebraic semantics of [4], Chapter 17, though a proper examination of this would be the topic for another paper.

https://doi.org/10.1017/S1755020318000473 Published online by Cambridge University Press

The arguments just given show that this detour through diagonalization is not necessary if ineffability is the goal, because (by Theorem 2.10 and Corollary 3.4) any such sporadic sentence over a given model will itself be ineffable; and as we have seen, there will always be such sporadic sentences. One upshot of the above analysis, then, is that ineffability really just is sporadicness.

§4. Concluding remarks. In [7] and [8], Welch has suggested that the existence of ineffable sentences may bear on the revenge immunity of Field's theory of truth. How might such a thought be made out? One natural idea would be that the ineffable sentences might constitute 'intuitively paradoxical' sentences that escape classification as defective at any level of the determinateness hierarchies available in the language of Field's theory.[20] If so, this would suggest an expressive weakness to that language, and threaten new revenge problems.

Of course a lot of the difficulty in framing the revenge problem here will be making sense of the idea that sentences like W are indeed 'intuitively paradoxical' in some meaningful sense of the term. I think a good case might be constructed for the claim that an ineffable sentence W should be considered intuitively paradoxical on Field's theory if at least one of the following two conditions are met:

1. There are *absolutely* ineffable sentences, in the sense that there are ineffable sentences that are ineffable over any ground model in the relevant language.
2. There is some paradoxical pattern of inference associated with such a W given a plausible set of rules and axioms 'read off' from Field's model theory.

On the first, the idea is that if ineffability is not absolute in this sense then it might seem plausible that the ineffability of a given sentence is really a reflection of necessary Tarskian constraints on given ground models for the language. For example, if ineffable sentences in 'smaller' models of set theory could be shown to be effable relative to 'bigger' ones, one would have some reason to believe that the ineffability phenomenon was somehow reflecting the restrictions imposed by the model constructions. On the other hand if ineffability is absolute in this sense one might be able to make a serious case that there really is a property of the sentence and its construction out of $\rightarrow$ and T that demands, but does not receive, treatment in the logic. To my knowledge, the question of the absoluteness of ineffability is open at present.[21]

On the second, it is helpful to compare the situation with the standard liar in the Kripke construction (from [5]), which is sometimes suggested to be an example of a revenge problem for a theory. Suppose we use Kripke's model theory to axiomatize a truth theory, taking as axioms the rules of strong Kleene logic, the rule

$$A/T^{\ulcorner}A^{\urcorner}$$

and its converse, together with a reasonably strong syntactic theory. Using these elements we can produce Q_0 and derive the rule

$$Q_0/\neg Q_0$$

and its converse. This is a particular sentence we have effective means to produce 'inside' the theory that exhibits what might naturally be construed as a paradoxical pattern of

[20] There's a notion of purely formal paradoxicality that might be identified with 'having value $\frac{1}{2}$ in a given model' that the ineffable sentences certainly have; but it is unclear what the philosophical significance of this is, for reasons that will become clear below.

[21] Welch has informed me that he has a sentence that works for all countable models; but the general case is still open.

https://doi.org/10.1017/S1755020318000473 Published online by Cambridge University Press

inference. Nevertheless, we know that any logic based on the Kripke theory must remain pretty much silent on the status of Q_0.[22] In particular, if the logic is encoded in a recursively enumerable validity relation $\vdash_K$, we will have $\nvdash_K Q_0$, $\nvdash_K \neg Q_0$, and so on. But then (the worry goes) surely this $\nvdash_K$ notion, which seems to be doing a lot of work in avoiding the nasty consequences of Q_0, should be expressible in the object language. And if we allow this, we might find ourselves saddled with paradoxes very similar to the ones we set out from.

Of course this is a very rough sketch and there are numerous points at which this case might be questioned. But *something* like this might be taken to be problematic for the Kripke theory; does W provide anything similar for Field's? Our foregoing observations suggest that there are considerable differences between W and Q_0: in particular, ineffable sentences are not primarily produced by diagonalization, and indeed the methods of argument presented do not provide any way of constructing an ineffable sentence W internally to the language of the theory, by means parallel to Gödel-Tarski diagonalization argument. (It would be a mistake to believe that Welch's $Q_W \equiv \neg D^W Q_W$ was a counterexample; for actually producing such a sentence Q_W requires producing a sentence W that is sporadic, but the arguments above show that this is really no advance on the problem of producing an ineffable sentence.) So it is somewhat unclear whether or not we can even produce an ineffable sentence step-by-step using Ts and $\rightarrow$s, and indeed whether we could in principle identify one on that basis if we saw one. Perhaps the only way of identifying them is in terms of the set-theoretic metatheory.

Even if we did find some ingenious way of combining Ts and $\rightarrow$s to produce an ineffable sentence, and even if that sentence was absolutely ineffable for set-theoretic models of the kind described above, this *still* would not settle the question of whether there is a distinctively paradoxical pattern of inference associated with W over some 'core rules' for Field's theory, comparable to that for Q_0. After all, what's valid on the standard model theory above is extraordinarily complex, and so any recursively enumerable validity relation $\vdash_F$ 'read off' from that model theory will be weaker. It is therefore possible that such a $\vdash_F$ would be consistent with accepting W and/or its negation as an axiom. If this could be shown (by modeling the chosen 'core rules' and an ineffable W), then we would have that W is consistent with the chosen 'core rules', and so (plausibly) not intuitively paradoxical from the point of view of the Field theory. In contrast, Q_0 cannot be accepted as an axiom on the Kripke theory without violating basic rules of the logic, basically because $Q_0 \vdash_K \neg Q_0$ on any reasonable axiomatization for $\vdash_K$.

To conclude, I would suggest that settling the two questions raised above, and especially the second, is an important open problem for assessing the revenge immunity of Field's theory of truth.

§5. Acknowledgments. I am extremely grateful to Hartry Field and Philip Welch for their helpful comments and encouragement. I am also greatly indebted to Sergei Artemov and Lavinia Picollo for their help on some technical points, to Neil Barton for useful comments on an earlier draft, and to an audience at the 'Semantic Paradox and Revenge' conference at the University of Salzburg in Summer of 2018, organized by Julien Murzi and Lorenzo Rossi, for helpful discussion.

[22] This applies whether we take strong Kleene or Supervaluation logic, though the situation is certainly starker when it comes to strong Kleene.

https://doi.org/10.1017/S1755020318000473 Published online by Cambridge University Press

BIBLIOGRAPHY

[1] Barwise, J. (1975). *Admissible Sets and Structures*: *An Approach to Definability Theory*. Berlin: Springer-Verlag.

[2] Field, H. (2003). A revenge-immune solution to the semantic paradoxes. *Journal of Philosophical Logic*, **32**(2), 139–177.

[3] Field, H. (2007). Solving the paradoxes, escaping revenge. In Beall, J. C., editor. *Revenge of the Liar*: *New Essays on the Paradox*. Oxford: Oxford University Press, pp. 53–144.

[4] Field, H. (2008). *Saving Truth from Paradox*. Oxford University Press.

[5] Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy*, **72**(19), 690–716.

[6] Moschovakis, Y. (1974). *Elementary Induction on Abstract Structures*. Dover Publications.

[7] Welch, P. (2008). Ultimate truth vis a vis stable truth. *Review of Symbolic Logic*, **1**(1), 126–142.

[8] Welch, P. (2011). Truth, logical validity and determinateness: A commentary on saving truth from paradox. *Review of Symbolic Logic*, **4**(3), 348–359.

[9] Welch, P. (2014). Some observations on truth hierarchies. *Review of Symbolic Logic*, **7**(1), 1–30.

[10] Welch, P. (2019). Some observations on truth hierarchies: A correction. *Review of Symbolic Logic*, doi:10.1017/S1755020319000042.

DEPARTMENT OF PHILOSOPHY
NEW YORK UNIVERSITY
NEW YORK CITY, NY 10003, USA
E-mail: cscambler@gmail.com

https://doi.org/10.1017/S1755020318000473 Published online by Cambridge University Press