More on Hyper-Reliability and A Priority[*]

James Pryor
NYU
<jim.pryor@nyu.edu>
9/25/06

Elaborates and corrects some things I said in Pryor 2006a.
The numbering of sections and examples continues that of the earlier paper.

VIII

In section III of Pryor 2006a, I argued against the view that *the mere fact that a thought-type is hyper-reliable* directly gives one justification to believe a thought of that type. A close alternative says that *our merely appreciating that* the thought-type is hyper-reliable directly gives us that justification.

We needed to refine these proposals to give them the best run for their money.

I gave examples of attitudes that would be self-verifying if one formed them, but where intuitively, that gave one no reason to form them. In response, the hyper-reliabilist can say he's only offering an account of doxastic justification—what justifies *a belief you have*—rather than an account of what gives you justification *to* believe some thought you don't yet have.

Another problem comes from examples of unknown (or empirically known) mathematical truths. Those truths will be true whenever entertained, but we don't think *that* by itself provides subjects with reason to believe them. The hyper-reliabilist can address this problem by saying it's only a certain *special kind* of hyper-reliability that gives the direct justification he's postulating. It's only when the subject's thinking a thought *is itself a truth-maker* for the thought that a subject has that direct justification. The mathematical thoughts would have been true anyway.

In Pryor 2006a, I presented a final example which was supposed to carry the day against even the most refined of these proposals. That example went like this:

> Consider this sentence:
>
> (8) I am uttering a sentence.

[*] I'm indebted to some very helpful discussions at Vermont, at St Andrews, at the Aristotelian Society, and in the 2006 Mind & Language seminar at NYU. Special thanks are due to Ralph Wedgwood.

> Given what this sentence means, it follows that whenever it's used to think a thought, that thought is true. (I count rehearsing a sentence to yourself privately as a kind of utterance.) And anyone who understands the sentence is in a position to know this. However, suppose you do utter the sentence (either privately or aloud). What then justifies you in believing that you are uttering it, or any sentence? It can't be your understanding of the sentence. That would only justify you in having beliefs about what's true *whenever the sentence is uttered*. It doesn't help you determine when that condition is fulfilled. The natural thing to say is that what justifies you in believing you are uttering the sentence is your introspective or perceptual awareness of uttering it. Hence, your justification for believing the thought you have by rehearsing (8) is a posteriori—despite the fact that you know, just by virtue of understanding (8), that whenever it's used to think a thought, that thought is true. (p. 334)

I stand by what I said there, but I now realize it's not the most *effective* example I might give. My opponent can protest that, since I count private rehearsals as utterances, my example in effect just reduces to:

(2) I am occurrently thinking.

And anyone who wants to count cogito judgments as justified just by the fact that one's thinking them makes them true will surely want to treat (2) this way too.

Now, that is not how I intended the example. Occurrent thinking is *not* always a matter of rehearsing sentences to oneself. Rather, (8) is a claim about *the particular way* in which one's occurrent thinking is unfolding. In my view, our justification about such matters comes from our experience or introspective awareness of doing the relevant thinking. It's on a par with "I am having a headache" and "I am now thinking of a prime number." I count such justification as, in some broad sense, "experiential" or a posteriori.

However, I acknowledge the worry that (8) may be too close to (2) to give us much substantial new leverage on the debate.

I now have some more effective examples, which present more starkly the kind of consideration I meant to be appealing to with (8). These examples consist of thoughts that are *made* true by one's thinking them, but where *recognizing* that fact requires considerable reflection and the application of logical/linguistic insight. For example:

(8*) Jeff and Mark discussed poetry last night, or I am having some thought that exemplifies non-distributive predication.

(8**) I am thinking some thought that includes quantification, and more than two unsaturated components.

Anyone who thinks these thoughts will need already to have the concepts of non-distributive predication, quantification, and unsaturated thought components; and they will, in thinking the

thoughts, make the thoughts true. But *ascertaining* that requires reflection. You have to *recognize that* these thoughts possess the very features they attribute—and that they're thoughts you yourself have.

I don't have any satisfying general account of what that recognition will consist in. But I think it's clear that it will rely on some amount of reasoning. That is, your justification for (8*) and (8**) won't be *direct* and non-inferential, the way the hyper-reliabilist claims that *cogito*-type thoughts like "I exist" and "I am occurrently thinking" are directly justified. (If you think you *can* directly tell that (8*) and (8**) are true, I invite you to substitute more complicated examples.) It will instead be inferential. That highlights the fact that your justification has two components: one component concerning the thought's logical properties, and the other concerning the fact that you are thinking it. The mere fact that thinking (8*) and (8**) makes them true doesn't *by itself* give subjects who believe them direct justification for their belief.

Now, if that's the right thing to say about (8*) and (8**), then it's natural to think that (1) "I exist" and (2) "I am occurrently thinking" have the same epistemic status. Your justification for these thoughts, too, has two components: a logical/linguistic recognition that thinking them makes them true, and an introspective awareness that they're thoughts now being thought by you. In the case of these thoughts, the logical/linguistic recognition is more easily achieved. Perhaps it even takes no reasoning. But two components are still required. Even when your *logical/linguistic* justification is direct and non-inferential, it still needs to be *combined with* an experience or introspective awareness that you are thinking the relevant thoughts. And so your justification still remains, in part, a posteriori. At best, (1) and (2) might be limiting cases of the epistemic situation we're in with (8*) and (8**): cases where the logical insight required is minimal. It's not plausible that (1) and (2) should get to be *directly and a priori justified* because of their self-truth-making character, when (8*) and (8**) do not.

Let's consider some responses.

One line of response is to a bite the bullet about (8*) and (8**). One can go "externalist" about our justification to believe them: that is, one can say that anyone who thinks those thoughts *does* thereby have justification to believe them, even if they don't yet "see it." "Seeing it" may be a requirement for properly *basing* a belief on this justification. But the justification *to* believe (8*) and (8**) is present and available wherever those thoughts are entertained.

That is an alternative worth considering seriously. In the end, I think the kind of "externalism" mooted here will be indigestible. I think it'll pressure us to say we already have justification to believe *all sorts of* still-unproven mathematical truths; and I think that's a wrong result. I have no conclusive refutation of the view. But I expect many of my readers will join me in thinking that dull-witted thinkers of (8*) and (8**) aren't yet justified in believing what they think. (It helps to construe "thinking" here to mean merely entertaining, rather than believing. It's somewhat more difficult to imagine subjects *believing* (8*) and (8**) without yet having "seen" that they make the thoughts true by thinking them; though arguably that is possible, too.)

A second response to my argument is to reject the analogy between (8*)/(8**) on the one hand, and (1)/(2) on the other. My opponents may say: "(8*)/(8**) are different because their *cogito*-like character is not logically transparent. (1)/(2) on the other hand *are* transparent. Anyone who properly understands (1)/(2) will see that thinking them make them true. So let us offer this new proposal: when a thought is such that the mere thinking of it makes it true, *and* the thought is logically simple enough, then someone who believes the thought thereby has a direct justification for their belief."

This sounds *ad hoc* to me. Why should properties like "transparency" and "logical simplicity" affect a thought's epistemic status—unless, as I'm claiming, that epistemic status owes in part to the subject's exercise of reflective and introspective insight?

But perhaps my opponent can find things to say here.

He may want to appeal to a suggestion that Ralph Wedgwood put to me in conversation. Wedgwood starts with the idea that some cognitive transitions are rational that *don't* just consist in the subject's thinking through an argument. For example, it may sometimes be rational to move from an experience as of red to a thought that there is something red before one. That's *not* a matter of moving from a belief or awareness *with the content* "I have an experience of red" to a thought about one's environment. Rather, one moves from the as-of-red experience *itself* to the thought about one's environment. The rationality of this transition *doesn't* derive from the states' contents lining up into a good argument. Even when they do so line up, that's not (all of) what *makes* the transition rational for the subject.

I embrace the view that cognitive transitions of this sort can be rational (see esp. my 2005). And even if I didn't, it's at least *intelligible* that there be rational transitions of the sort Wedgwood is describing.

Wedgwood deploys this idea in defense of my opponent. Perhaps, he suggested, transitions from the thought "Thoughts of type T are true whenever they're thought" to particular thoughts *of* type T are intrinsically rational. The subject needn't recognize that *he is* thinking a thought of type T; he just needs to do it.

Of course, as Wedgwood acknowledged, we're owed a story about *why* so-and-so particular transitions should be rational ones. There's a promising candidate at hand: namely, transitions from the thought "Thoughts of type T are true whenever they're thought" to particular thoughts of type T are *guaranteed to be truth-preserving*. If the first thought is true, then the second one must be true as well. Introspection plays no role in justifying this transition. So if (as is plausible) a subject can have a priori justification for the first thought—that cogito thoughts are true whenever they're had—then the particular cogito thought he transitions to ought to retain that a priority.

This is an elegant proposal that deserves careful thought. Again I'm in no position to offer a conclusive refutation. But I will voice a worry.

Suppose a subject occurrently believes, and has a priori justification for believing, that 65+78=143. Now consider the thought (that is, the act of entertaining the hypothesis) *I believe that 65+78=143*. Intuitively, this thought should get what justification it has from your experience or introspective awareness of your occurrent mental life. So its justification will be a posteriori.

Notice that *I believe that 65+78=143* is not a cogito thought. Thinking it doesn't make it true. You can think you believe things that you don't thereby believe. So when I invite you to agree that this thought is, intuitively, justified a posteriori, I'm not begging any questions about the epistemology of the cogito.

But now consider *the transition* from the belief that 65+78=143 to the thought that you have that belief. Just as Wedgwood observed in the case of cogito thoughts, this transition too is guaranteed to be truth-preserving. Anyone who transitions from the belief that P to the thought *I believe that P* (without thereby abandoning the starting belief) ends up with a true thought. If this guaranteed truth-preservation was enough to make the cogito transition non-introspectively justified, then it ought to do so here, too. And, continuing to carry over Wedgwood's analysis, since the state you started with (the belief that 65+78=143) is one you have a priori justification for, the state you transition *to* ought again to retain that a priority. But this seems to be a wrong

result. Intuitively, your justification to believe *I believe that 65+78=143 is* a posteriori, even if (as is plausible) you formed that thought directly in response to your first-order belief that 65+78=143. I think something's awry with any analysis that says *I believe that 65+78=143* is justified a priori.

There is more to be thought through here. Perhaps there are natural disanalogies between the cogito case and the math belief case, that can explain why Wedgwood's proposal about the cogito shouldn't carry across. Or perhaps the conclusion that *I believe that 65+78=143 is* justified a priori is one we should reconcile ourselves to, after all. But in my view, the most plausible analysis is that *all* of these thoughts owe their justification in part to introspective experience.

IX

I'm now dissatisfied with the exact details of the free logic I was employing in sections IV-V of Pryor 2006a. The larger picture I was arguing for there should remain standing; but I want also to get the details right.

In the earlier paper, I opted for a neutral free logic with weak Kleene connectives and bivalent quantifiers. (In doing so, I followed Lehmann 1994.) I now think my impression that that particular free logic was especially suited to my purposes was confused; and moreover, I think that a different particular free logic is better motivated.

First, let me review and elaborate the larger picture that I still subscribe to.

I want to distinguish between two kinds of "entailments" that a thought can have. On the one hand there are *properly logical* entailments, which are in some vague sense "internal" to the contents that our thoughts actually have. On the other hand there are things that have to be the case *for the thoughts to have* those contents. I'll call the former **proper entailments**, and the latter **meta-cognitive conditions**. For example, a proper entailment of `Gα & Gβ` would be `Gα`. The following are instead meta-cognitive conditions: `there are conjunctive thoughts`, `there are thoughts some of whose components are unsaturated`, and `there is such a proposition as Gα & Gβ`. It will be useful to speak of these collectively as "entailments" of a thought, though in my view the meta-cognitive conditions aren't really any kind of entailment of the thought itself. They're entailments of meta-cognitive facts *about* the thought.

These different kinds of "entailment" may have different epistemic properties. In particular, I hope to articulate the *logical* relations among our thoughts in such a way that *properly logical* entailments preserve a priority. Meta-cognitive conditions are not bound by that same constraint. For example, since it's not a properly logical entailment of `Gα & Gβ` that `there is such a proposition as Gα & Gβ`, the latter need not be *logically* and a priori inferable from the former. (In this particular example, the latter is plausibly a priori knowable on *independent*, philosophical grounds.)

I want us to use a free logic which can distinguish between existentially "hedged" and existentially unhedged predications. We'll understand hedging in terms of its semantic power, rather than in terms of any specific logical form. An unhedged predication of `G` to α is one that entails `α exists`; a hedged predication is one that does not. A thought that predicates `G` to α in the hedged way won't *logically* entail `α exists`; and so, in the framework I'm setting out, the latter need not through *logic* be a priori inferable from the former. It may at the same time be true that α is a McDowell/Evans-style demonstrative, and so the thought's content is only available to be thought when α refers. In such a case, `α exists` would be a meta-cognitive condition for our hedged thought, without being a proper entailment. Were the demonstrative α not to refer, the hedged thought and its unhedged counterpart would *both* be contentless. But only the unhedged thought will properly entail, and so permit us to infer a priori, that `α exists`. The crucial point here is that **the meta-cognitive conditions of our thoughts need not be a priori consequences of them**. To cognitively avail oneself of those meta-cognitive conditions, one will normally need to *establish that one has* the thoughts in question: which will already take one into the realm of the a posteriori.

The details of all this will depend on what we say are the "properly logical" relations among our thoughts. Let's think this through first at the level of the sentences through rehearsing which we think our thoughts. We'll be talking about logical relations these sentences have to each other, and also to sentences containing non-referring terms—the latter of which not being thinkable, on the McDowell/Evans-style view. Later we'll consider whether our analysis can be carried over to our contentful thoughts *themselves*, bypassing any sentential intermediaries.

So we begin with the logic of sentences.

One way to secure a distinction between hedged and unhedged predication is with what's called a **positive free logic**: a free logic that allows atomic predications with non-referring terms

to sometimes be true. For example, a positive free logic may count `Pegasus = Pegasus` as true, even though `Pegasus` doesn't refer and `Pegasus exists` is false. More controversially, a positive free logic may count `Pegasus flies`, or (the non-atomic) `Pegasus eats meat ∨ ~Pegasus eats meat`, as true. For a positive free logic, `Pegasus flies` will already be hedged. To get an unhedged predication, we'd need something like `Pegasus exists and flies`.

I'm well-disposed towards positive free logics. Many philosophers are uncomfortable with them because they think they commit us to a Meinongian ontology. I don't share that fear. However, it's best not to fight more battles than you have to. So for this discussion, I'm not going to make use of any positive free logic.

The alternatives are **negative free logics**, where atomic predications with non-referring terms are always false, and **neutral free logics**, where atomic predications with non-referring terms (normally) take a third, undesignated truth-value (there may be exceptions for = and `exists`).

When writing Pryor 2006a, I thought that a neutral free logic best fit the McDowell/Evans view that demonstration failures result in failures to contentfully think. I was thinking that the third truth-value would attach to sentences that failed to be thinkable due to reference-failure. (Thus the weak Kleene connectives: truth-functional compounds of unthinkable sentences should also be unthinkable.) But then (following Lehmann) I went on to make quantifiers bivalent, so that `∃x: x = α` counted as false when α doesn't refer. That was needed to secure the right kind of difference between, e.g., a hedged predication of self-identity to α and an ordinary predication of self-identity. The hedged predication came out a theorem, and not entailing `α exists`; whereas the ordinary predication came out not a theorem, and entailing `α exists`.

It's an odd feature of that logic that `β = β ∨ α = α` comes out neither true nor false, with β referring but α not, yet `∃x: x = x` comes out true. ∃ is no longer a generalized form of ∨. But more importantly, I hadn't digested the tension between my analysis of hedging and what motivated me to choose a neutral free logic in the first place. I was assigning "true" to some thoughts that exhibited reference-failure: most notably, to hedged predications like `∀x: (x = α ⊃ x = x)`, where α is non-referring. But my initial motivations for choosing a neutral free logic should have precluded that.

So *I* was clearly confused. I should have recognized that conflict. I've now come to think that the initial motivations I was being guided by were also confused.

The role of our free semantics is to assign certain *marks* or *values* to formulas. The role of those values is to help us trace entailment relations. In doing this, we needn't require that the formulas in question be contentfully thinkable. It's enough that they have *a structure* that puts them into logical relations with other sentences, some of which *are* contentfully thinkable. An unthinkable sentence can still be assigned a value—perhaps it can even be assigned the value "T." Doing so can help us to better articulate some logical differences exemplified among *the thinkable* sentences: like the difference between hedged predications and ordinary predications to the same existing objects.

It may offend to call sentences that aren't contentfully thinkable *true*. Don't get hung up on that. I'll refer to our semantic values as "T," "F," and when there's a third value, "N." Don't think of "T" as meaning *true*. Think of it as meaning *truth-like*. A sentence only counts as *true* when it's both T *and* contentfully thinkable.

Here's a helpful analogy. You have a bunch of soap opera scripts that partition into the Trashy, the Farcical, and the Noxious. Only some of these scripts are producible. Noxious scripts are never producible; but neither are some Trashy and some Farcical scripts. These scripts stand in certain indebtedness relations to each other (they share characters, have plot crossovers, spoof each other, and so on). We may be able to *identify* different patterns in the borrowing relations among our produced scripts, yet those patterns only be *intelligible* when we take into account relations across *the whole range* of scripts, both produced and unproducible. In the same way, I claim, the different patterns of proper logical entailment among thinkable sentences will only be intelligible when we take into account structural relations they stand in to unthinkable sentences, too.

If we think of the semantic project in the way I've described, then the mere fact that demonstration failures preclude contentful thinking doesn't settle the question which free semantics and logic we should employ. On what grounds then *should* we choose a (non-positive) free logic?

Here's my current thinking.

In crafting a device for hedged predication, we ought to recognize the possibility of "partial hedging." For example, consider:

(9) Jack is younger than Jiho.

On any classical, or non-positive free logic, this is an unhedged claim that entails that both Jack and Jiho exist. A *partial* hedge would look something like this:

(9*) (If Jack exists) Jack is younger than Jiho.

Leave aside the question of how exactly to implement that "hedge." The intuitive idea is that such a partially hedged claim ought *not* to entail that Jack exists, but *ought still* to entail that Jiho exists.

Now, if we think through the various natural choices for a non-positive free logic, only one turns out to straightforwardly give us a form of hedging with that result.

I'll understand $\Gamma \models \Delta$ to require that there is no model M on which every sentence in Γ is T but Δ is non-T.[1]

On the strictest "Fregean" free logics, which make any formula containing a non-referring term neither true nor false, there can't be *any* kind of hedged predication (If α exists) φ that fails to entail α exists. Every model on which α refers will make the conclusion α exists T, and every model on which α fails to refer must make the premise N. So the premise will entail the conclusion.

On a negative free logic, like the one in Burge 1974, atomic predications $G\alpha$ are *false* when α doesn't refer, as are claims like α exists. So one can hedge a claim φ against α's nonexistence with α exists $\supset \varphi$. Whenever α fails to exist, the antecedent will be F and hence the conditional will be T. The problem is that this kind of hedge is all-or-nothing. We can't *partially* hedge a claim, against only *some objects'* nonexistence. For instance, the claim:

(9*⊃) Jack exists ⊃ Jack is younger than Jiho.

will also be T whenever Jack fails to refer—even if Jiho also fails to refer. So (9*⊃) won't entail Jiho exists, as we think a partial hedge intuitively should.

---

[1] One might choose to require, *additionally*, that there is no model M on which every sentence in Γ is non-F but Δ is F. The stronger construal of $\models$ can be imposed on some of the free logics I review below, but not all of them. It isn't easily imposed on the free logic that gives us the best account of "hedging."

On a neutral free logic like the one in Lehmann 1994, and my 2006a, we need to bring in quantifiers to craft our hedged claims. The most straightforward formula would be this (an equivalent of what I offered in my earlier paper):

(9*∀) `∀x: (Jack = x ⊃ x is younger than Jiho).`

where the value of `∀x:` φ on a model M and assignment V, written $|\forall x\text{:}\ \varphi|^{M}_{V}$, comes out F if there's any assignment V* (differing from V at most with respect to what it assigns to `x`) such that $|\varphi|^{M}_{V^*}$ is F; otherwise $|\forall x\text{:}\ \varphi|^{M}_{V}$ is T. That will do fine so long as `Jiho` refers. In that case, if `Jack` also refers, (9*∀) will be T whenever `Jack is younger than Jiho` is T. And if `Jack` doesn't refer, then the antecedent of `Jack = x ⊃ x is younger than Jiho` will be N on every assignment, and so the whole conditional will never be F, and hence (9*∀) will again be T. So far, so good. The problem arises when `Jiho` doesn't refer. Intuitively, since (9*∀) is only partially hedged, it should still entail `Jiho exists`. So it shouldn't ever be that (9*∀) is T when `Jiho exists` is not. Unfortunately, on the semantics we've here specified, that's exactly what does happen, when `Jiho` fails to refer.

I think the best account of hedging, including partial hedging, is got from a neutral free logic that uses *strong* Kleene connectives and *tri*valent quantifiers. On such a logic, if we understand F < N < T, then $|\varphi\ \&\ \psi|^{M}_{V}$ will be min($|\varphi|^{M}_{V}$, $|\psi|^{M}_{V}$); and $|\forall x\text{:}\ \varphi|^{M}_{V}$ will be min($|\varphi|^{M}_{V^*}$ for each V* differing from V at most wrt `x`), or T when M's domain is empty. There are two options for how to construe $|\texttt{Jack = x}|^{M}_{V}$ when `Jack` doesn't refer on M. According to one, $|\texttt{Jack = x}|^{M}_{V}$ is F for every V. According to the other, it's always N.

If we make $|\texttt{Jack = x}|^{M}_{V}$ always F, then the antecedent of `Jack = x ⊃ x is younger than Jiho` will be F on every assignment, and the whole conditional will therefore be T—regardless of whether `Jiho` refers on M. So (9*∀) will again be T even when `Jiho exists` is not T. (9*∀) again fails to give us the entailment to `Jiho exists` that we think a partial hedge should give us.

We only get a good account of partial hedging by taking the second option, of making $|\texttt{Jack = x}|^{M}_{V}$ always N when `Jack` doesn't refer. Then when `Jack` doesn't refer, the antecedent of `Jack = x ⊃ x is younger than Jiho` will always be N; if `Jiho` doesn't refer either, so too

the consequent; making the whole conditional N on every assignment. So (9*∀) is now N, not T. We're no longer in a situation where (9*∀) is T when `Jiho exists` is not T. So on this interpretation, (9*∀) *can* finally entail `Jiho exists`.

As we'll see below, on this logic, a hedged predication `∀x: (α = x ⊃ φ)`, with α not occurring in φ, will never entail `α exists`. It may entail `β exists` for other terms β occurring in φ. The hedged predication will be a theorem when and only when `α exists` entails `φ[α/x]`, the result of replacing every free occurrence of `x` in φ with α. I think this is the non-positive free logic best suited for crafting a device for hedged predication.

This logic does have the somewhat surprising feature that, with β referring but α not, α=β comes out not F but N. You might intuit to the contrary: that it is genuinely *false* that, say, Pegasus is identical to you.

However, in the first place, any awkwardness here should be offset by the fact that this is the only natural logic that properly handles partial hedging.

In the second place, I don't think we should be *having* direct intuitions about which truth-like value T, F, or N, our semantics should assign. Consider an analogy. Montague semantics that treat names as generalized quantifiers will say the semantic value of "Jack" is the set of all predicate extensions containing some individual. That has certain theoretical benefits. It may or may not ultimately be the best semantic theory; but I assume it's *no* good objection to it that "Intuitively, 'Jack' names a person not a set of sets!" The Montague semantics is not (directly) proposing any analysis of our folk relation of "naming." It should be judged by its final theoretical predictions—not by our ability to find folksy correlates for internal pieces of its machinery.

So too in our case. Our assignments of T, F, and N are part of a theoretical apparatus; they're not meant to (directly) capture our folk notions of "true" and "false." The place where the theoretical apparatus should be judged against our intuitions is in its predictions about what sentences are designated and what entailments are valid. (Given a McDowell/Evans-style view about what sentences are thinkable, this arena may perhaps be even further restricted, to which *contentfully thinkable* sentences are said to be designated, and what entailments between *contentfully thinkable* sentences are said to be valid.) Direct judgments about which sentences should come out N, which come out F, and so on, have no intuitive authority.

X

Here are **semantics** for the neutral free logic I'm endorsing.

A model M for our language is a pair of an interpretation function $|\bullet|^M_\bullet$ and a (possibly empty) domain. An assignment V on M is a total function from the language's variables to M's domain.

Our "truth-like values" are T, F, and N. T is the only designated value. $\Gamma \models A$ means that there's no model on which every sentence in Γ is designated but A is undesignated. $\models A$ means that A is designated on every model.

If x is a variable, then $|x|^M_V$ will be $V(x)$. Names and atomic predicates have the same interpretation relative to every assignment (which we can write as $|\bullet|^M$, leaving off the V). The interpretation of a name is either some object from M's domain, or undefined. The interpretation of an n-place atomic predicate is a function from n-tuples of objects in M's domain to {T,F}.

Let $t_1 \ldots t_n$ be arbitrary terms of the language (either names or variables), and G an n-place atomic predicate. Then $|Gt_1 \ldots t_n|^M_V$ will be N if any of the $|t_i|^M_V$ are undefined, otherwise it will be the value of $|G|^M_V$ for $\langle |t_1|^M_V \ldots |t_n|^M_V \rangle$.

$|t_1 = t_2|^M_V$ will be N if either $|t_1|^M_V$ or $|t_2|^M_V$ are undefined, otherwise it will be T if $|t_1|^M_V = |t_2|^M_V$, otherwise it will be F.

$|\varphi \vee \psi|^M_V$ will be $\max(|\varphi|^M_V, |\psi|^M_V)$, understanding F < N < T. $|\exists x\colon \varphi|^M_V$ will be F when M's domain is empty, otherwise $\max(|\varphi|^M_{V^*}$ for each V* differing from V at most wrt x).

$|\varphi \,\&\, \psi|^M_V$ will be $\min(|\varphi|^M_V, |\psi|^M_V)$. $|\forall x\colon \varphi|^M_V$ will be T when M's domain is empty, otherwise $\min(|\varphi|^M_{V^*}$ for each V* differing from V at most wrt x).

$|\sim\varphi|^M_V$ will be F if $|\varphi|^M_V$ is T, T if $|\varphi|^M_V$ is F, and N otherwise.

$\varphi \supset \psi$ will be semantically equivalent to $\sim\varphi \vee \psi$.

t exists will be semantically equivalent to $\exists x\colon t = x$. It will be T when t is defined, and otherwise N.

A few remarks about the correlative **proof theory**. In the following, α and β are names; A, B, and C are closed sentences; X and Y are arbitrary sets of closed sentences; φ is a formula open at most in x; and φ[α/x] is the result of replacing every free occurrence of x in φ with α. Many familiar rules will be valid:

v+ if X ⊢ A, then X ⊢ A ∨ B and X ⊢ B ∨ A.

v– if X, A ⊢ C and Y, B ⊢ C, then X, Y, A ∨ B ⊢ C.

&+ if X ⊢ A and Y ⊢ B, then X, Y ⊢ A & B.

&– if X, A ⊢ C, then X, A & B ⊢ C and X, B & A ⊢ C.

dbl neg+ if X ⊢ A then X ⊢ ~~A.

modus ponens if X ⊢ A and Y, B ⊢ C, then X, Y, A ⊃ B ⊢ C.

The logic will be in some respects *stronger* than minimal and intuitionistic logic: for example, it will include **ex falso** (which minimal logic does not include) and **double negation elimination** (which neither includes).

ex falso if X ⊢ ~A then X, A ⊢ C.

dbl neg– if X, A ⊢ C then X, ~~A ⊢ C.

It will also include the following rule, which is valid in classical logic but not in minimal or intuitionistic logic:

if X ⊢ A ⊃ C then X ⊢ ~A ∨ C.

On the other hand, none of these three logics includes excluded middle. And in other respects, the neutral free logic will be *weaker* than minimal and intuitionistic logic. For example, none of the following rules will be valid:

reductio if X, A ⊢ C and X, A ⊢ ~C then X ⊢ ~A.

cp if X, A ⊢ C then X ⊢ A ⊃ C.

if X, A ⊢ C then X ⊢ ~A ∨ C.

if X, A ⊢ C and X ⊢ C ⊃ B then X ⊢ A ⊃ B.

if X ⊢ A ∨ D then X ⊢ (D ⊃ A) ⊃ A.

These rules will be violated, for instance, when X permits A and B to be N, C to be F, and D to be T. *Restricted* forms of those rules will be valid in our neutral free logic. For example:

weak cp if X, A ⊢ C and X ⊢ A ∨ ~A then X ⊢ A ⊃ C.

weakening if X ⊢ C then X ⊢ A ⊃ C.

Let's consider the predicational part of our free logic.

∀+ if X, α exists ⊢ φ[α/x], with α not occurring in X or φ, then X ⊢ ∀x: φ.

∀– if X, φ[α/x] ⊢ A, and Y ⊢ α exists, then X, Y, ∀x: φ ⊢ A.

∃+ if X ⊢ φ[α/x], and Y ⊢ α exists, then X, Y ⊢ ∃x: φ.

∃– if X, α exists, φ[α/x] ⊢ A, with α not occurring in X, φ, or A, then X, ∃x: φ ⊢ A.

=+ if X ⊢ α exists, then X ⊢ α = α.

=– if X ⊢ φ[α/x], then X, α = β ⊢ φ[β/x] and X, β = α ⊢ φ[β/x].

existence if φ is an elementary formula containing α (that is, φ is atomic, or an identity, or the negation of either), then φ ⊢ α exists

Soundness (and independence) proofs for these rules are straightforward. I'll give some examples. They rely on the following Lemmas, which are readily proved by induction on the complexity of formulas:

Lemma 1. If α doesn't occur in φ, and M* is a model that differs from M at most wrt what it assigns to α, then $|\varphi|^{M}_{V}$ will = $|\varphi|^{M^*}_{V}$.

Lemma 2. If M is a model that assigns o to α, and V is an assignment that assigns o to x, then $|\varphi|^{M}_{V}$ will = $|\varphi[\alpha/x]|^{M}$.

Sample soundness proofs

Proof of **∨+**. If a model M is to make X T but A ∨ B non-T, then by the semantics for ∨ it needs to make A non-T. (Similarly for B ∨ A.) But then M makes X T but A non-T, contrary to the hypothesis that X ⊨ A.

Proof of **∨–**. If a model M is to make X, Y, and A ∨ B T but C non-T, then by the semantics for ∨ it needs to either (i) make A T, in which case M makes X, A both T but C non-T, contrary to the hypothesis that X, A ⊨ C; or (i) make B T, in which case M makes Y, B both T but C non-T, contrary to the hypothesis that Y, B ⊨ C.

Proof of **∀+**. If a model M is to make X T but ∀x: φ non-T, there must be some object o in M's domain, and assignment V that assigns o to x, such that $|\varphi|^{M}_{V}$ is non-T. Let M* be a model that differs from M by assigning o to α. Since α doesn't occur in φ, Lemma 1 tells us that the non-T $|\varphi|^{M}_{V}$ will = $|\varphi|^{M^*}_{V}$, which Lemma 2 tells us will = $|\varphi[\alpha/x]|^{M^*}$. But $|\alpha\ \texttt{exists}|^{M^*}$ will be T, and since α doesn't occur in X, Lemma 1 tells us that the T $|x|^{M}_{V}$ will = $|x|^{M^*}_{V}$. But now M* makes X and α exists T but φ[α/x] non-T, contrary to the hypothesis that X, α exists ⊨ φ[α/x].

Proof of **∀–**. If a model M is to make X, Y, and ∀x: φ T but A non-T, then since Y is T and Y ⊨ α exists, α exists must also be T, so M must assign some object o to α. For M to make ∀x: φ T, since M has at least o in its domain, there must be an assignment V that assigns o to x, and $|\varphi|^{M}_{V}$ must be T. By Lemma 2, the T $|\varphi|^{M}_{V}$ will = $|\varphi[\alpha/x]|^{M}_{V}$. But now M makes X and φ[α/x] T but A non-T, contrary to the hypothesis that X, φ[α/x] ⊨ A.

The rules I've set out are all sound on the semantics I specified. I think they don't yet constitute a complete proof theory. That doesn't matter for our purposes.

I said at the end of section IX that **a hedged predication ∀x: (α = x ⊃ φ), with α not occurring in φ, will never entail α exists; and will be a theorem just in case α exists entails φ[α/x]**. Here are the proofs.

First: let M be any model assigning nothing to α, but where every object in its domain satisfies φ (if φ is something like x≠x, this requires giving M an empty domain). Then $|\alpha = x \supset \varphi|^{M}_{V}$ will be T wrt every assignment V, and so $|\forall x: (\alpha = x \supset \varphi)|^{M}$ will be T, but $|\alpha\ \texttt{exists}|^{M}$ non-T. So ∀x: (α = x ⊃ φ) cannot entail α exists.

Second: suppose that Υ, α exists ⊢ φ[α/x], with α not in Υ or φ. Then by **∀+**, Υ ⊢ ∀x: φ. Choose a β not in Υ or φ. Then by **∀-**, Υ, β exists ⊢ φ[β/x]. Then by **∨+**, Υ, β exists ⊢ φ[β/x] ∨ α≠x. Using a variant of **weak cp** we can translate that to Υ, β exists ⊢ α=x ⊃ φ[β/x]; and with **∀+**, we get Υ ⊢ ∀x: (α=x ⊃ φ). In conclusion, then, when Υ, α exists ⊢ φ[α/x], it follows that Υ ⊢ ∀x: (α=x ⊃ φ). Setting Υ to ∅, we get that ∀x: (α = x ⊃ φ)is a theorem whenever α exists entails φ[α/x].

In the opposite direction: from α exists, rule **=+** permits us to derive α = α, and from ∀x: (α = x ⊃ φ) and α exists, rule **∀–** permits us to derive α = α ⊃ φ[α/x]. From there, **∨–** and **ex falso** permit us to derive φ[α/x]. So whenever Υ ⊢ ∀x: (α = x ⊃ φ), it will follow that Υ, α exists ⊢ φ[α/x]. (It's worth noting a further consequence: using **∀+**, we can derive Υ ⊢ ∀x: φ.) Setting Υ again to ∅, we get that whenever ∀x: (α = x ⊃ φ)is a theorem, α exists entails φ[α/x]. (And also as a theorem that Υ ⊢ ∀x: φ.)

A not-particularly welcome feature of the semantics is that there won't be any models in which ~α exists is true; and hence ~α exists ⊨ C, for any C.

We might try to avoid that result by making "exists" a primitive logic predicate, such that when $|t|^{M}_{V}$ is undefined, $|t \text{ exists}|^{M}_{V}$ is F rather than N. But that doesn't really dispel the problem. It'd still be true that ~(∃x: x = α) ⊨ C, for any C.

I admit this is an uncomfortable result. It should be weighed against this being the only natural semantics that properly handles partial hedging. Additionally, the problem is confined to sentences like ~(∃x: x = α), with non-referring α, that by McDowell and Evans' lights aren't thinkable, anyway.

XI

We've so far been discussing the logic of sentences. But thinking is not always done with sentences. To what extent can we translate our proposal into an account of logical relations among our contentful thoughts, themselves? Might our thoughts *themselves* have enough structure in them to underwrite the logical relations we need?

That depends.

Consider two kinds of cognitive situation. In GOOD cases, you perceive an apple, successfully demonstrate it, and judge *That is a piece of fruit*. In BAD cases, you *seem* from the inside to be in a good case; but in fact there is no apple, your demonstration doesn't succeed, and according to McDowell and Evans you're not contentfully thinking any atomic thought that predicates *is a piece a fruit*. (You may manage to think various more general and *non*-demonstrative thoughts, such as *I see some piece of fruit*, *I am demonstrating an apple and predicating being a piece of fruit to it*, and so on.)

Intuitively there seems to be much in common to your cognitive situation in these different cases.

Consider first two good cases, where you're demonstrating and thinking about different apples. Here a McDowell/Evans-inspired theorist will deny that your atomic thoughts *That is a piece of fruit* have the same content. But he can allow that your thoughts are identically *structured*, and involve some of the same components: they both predicate the content *is a piece of fruit*. (Perhaps your thinking is in both cases a cognitive relation to an *abstract structured content*. Or perhaps the structure is only to be found in your cognitive activity itself. Different theories we'll be considering go different ways about this.)

How shall we explain the apparent commonality between good cases and bad cases? One way of capturing that is to say that your thought in a bad case *does* have a content, but it's a specially defective or "gappy" content that cannot possibly be true. (David Braun and others defend views like this.) Here again the cognitive commonality consists of some shared structure between your thoughts. The gappy thought and the good thought both involve predications of the content *is a piece of fruit*. A view of this sort can reproduce much of what McDowell and Evans wanted. But it doesn't conform to the letter of their view. On this gappy theory you in fact *do* have contentful thoughts in the bad cases.

We need a minor adjustment to bring the gappy theory into line with the letter of what McDowell and Evans said. We need to say that the cognitive episode you're undergoing in a bad case doesn't really count as *a contentful thinking*. We can still allow that it's an episode with a certain structure. Perhaps it's even a cognitive relation *to* a structured content-like thing. It's just that we refuse to call anything gappy a *genuine content*; and refuse to call these episodes *contentful thoughts*.

Such a view would respect much of the letter of McDowell and Evans' proposals. But it's probably not faithful to their intent. Their intent is probably better captured by a "disjunctive" picture, which denies that your relevant cognitive state in a bad case genuinely shares *any* structural properties with your states in the good cases.

We have three pictures on the table: straight gappy content theory, gappy content theory in McDowell/Evans clothing, and the disjunctive construal of McDowell and Evans. So far as I can see, there's no obstacle to carrying over our account of free logic for *sentences* to the *structured thoughts* or *contents* or *content-like things* postulated by the first two pictures. On the third picture, though, it looks like our thoughts don't yet give us a family of structures broad enough to apply the semantics we want. Our neutral free semantics requires the thinkable structures to share structure with *other* structures, that are neither T nor F, and so for McDowell and Evans wouldn't even be thinkable.

However, all is not lost.

Suppose Joe is a classical mathematician who works amongst non-classical logicians. He doesn't think the model theories his colleagues put forward are *true*; that is they do not correctly capture the semantic properties of his mathematical language. However, Joe is *interested in* the special classes of mathematical results that are provable in the ways his colleagues discuss. In his dealings with his colleagues' $\vdash$ relations, it's sometimes useful for him to make use of their model theories. But Joe keeps a critical distance. He says to himself, "This is not a correct account of the real meanings, or real truth-values, of my mathematical language. It's just a systematic distribution of some values. The $\vdash$ relation I'm dealing with is the one that's sound and complete with respect to preserving certain of those values. The model theory's value is in enabling me to identify and engage in metalogical reasoning about that $\vdash$ relation." I think Joe's attitude here is perfectly legitimate.

I want to urge the McDowell/Evans-inspired theorists to make the same kind of move. They should *find* a way to systematically associate sentences or other structures with their thoughts, structures that are part of *a larger family* of structures not all of which line up with cases where something is genuinely thinkable. We'll apply our free logical semantics to that family of structures. The McDowell/Evans-style theorist need not agree that the result accurately captures the real semantic properties of our thoughts. But they can allow that it determines a $\vdash$ relation. And nothing bars them from attaching special epistemic significance to that relation. In

particular, they're free to say that *the restriction of this* $\vdash$ *relation to the structures associated with genuinely contentful thoughts marks out a set of a priori justifiable inferences between our thoughts*. They're free to *deny* that the *real* semantic entailment relation between their thoughts has the same epistemic significance.

I said they're *free* to say all this. Is there any reason why they *should*?

Well, in my original 2006a paper, I noted that, pre-theoretically, it seems right to say that *experience* is part of what justifies us in believing:

(4) Jack exists,

and moreover that the degree to which it's rational believe (4) is less than 1. At the same time, it seems right to say that *something like*:

(5) Jack is self-identical

is justified just by *our understanding* of the notions of identity and so on, and so should be a priori. Moreover, here it *does* seem rational to believe the relevant claim to degree 1. This all constitutes a puzzle, since (5) classically entails (4). My aim in these two papers has been to get as close to those pre-theoretic thoughts as we can. I've done so by identifying a close surrogate for (5)—the hedged claim:

(5*$\forall$) $\forall$x: (x=Jack $\supset$ x is self-identical)

and a proof relation $\vdash$ on which that surrogate does not entail (4). This proof relation is the one determined by the free semantics I spelled out above.

The McDowell/Evans-inspired theorist *can* help himself to the claim that that proof relation marks what's a priori inferable from what, without having to agree that our thoughts really have the contents that my free semantics assigns them. That's analogous to the classical mathematician helping himself to the claim that, say, intuitionistic $\vdash$ really does mark what's constructively provable from what, without agreeing that his thoughts really have the semantics an intuitionistic model theory assigns them.

Is it especially *appealing* to divorce the proof relation that marks a priori inferability from the real semantic facts about our thoughts' contents, and what semantic entailment relations really hold among them? No, I don't think it is. But it is a move that's available.

In the dialectical setting we're now considering, three pressures are coming into conflict. There's:

(a) the desire to preserve our pre-theoretic judgments about (4) and (5).

I argued that, at the *sentential* level, this is best accomplished with the free semantics set out above. There's:

(b) the desire that the proof relation that marks a priori inferability be in synch with the real semantic entailment relation for our thoughts.

Finally, there's:

(c) the desire to keep to the true spirit of McDowell and Evans' theory of demonstrative thought.

(c) has the result that there aren't structural commonalities between our thoughts in good and bad cases. This prevents us from applying our free semantics to the thoughts' own contents.

One of these has to give. I urge that it not be (a).

XII

Let me rehearse again the main lessons I drew in 2006a. On the view I'm proposing, experiences of Jack are necessary to be able to have *any* of the following thoughts:

(4) Jack exists.

(5) Jack is self-identical.

(5*∀) ∀x: (x=Jack ⊃ x=x)

In the first two cases, these experiences play a justifying role. In the last case, they do not. You can know a priori that *each* of the thought-types is hyper-reliable, and can only be successfully had when Jack exists. But that doesn't make *any particular thoughts* of those types justified a priori. In cases where you manage to successfully think a (5*∀) thought, that thought *is* justified a priori, through your understanding of the logical relations it involves. But it doesn't logically or a priori entail (4) or (5).

Your a priori justification to believe instances of (5*∀) sits alongside a *lack* of justification to believe *you are* successfully thinking such thoughts, and so satisfy their meta-cognitive conditions. From the premise *that* you're thinking a (5*∀) thought, you *could*, on a McDowell/Evans view, infer a priori that Jack exists. But from the particular (5*∀) thought

itself, you cannot. It takes more to justify you in believing *you have* (5*∀) than it takes to justify the thought itself.

When it's an open question for you whether Jack exists, it may seem peculiar to try to think thoughts like (5*∀), without yet knowing that you'll succeed. (It's somewhat like prefacing an email with "If you're still reading email at this address …" That insures you no better against not being read.) However, you don't really have better alternatives. If *agnosticism* is assigning a middling degree of belief to a thought, then your ability to have *that* kind of attitude towards (5*∀) is just as threatened by the prospect of Jack's not existing as full belief towards (5*∀) would be. There is such a condition as *having no degree of belief at all* towards a thought: e.g., if you've never considered the thought, or are incapable of considering it. But when you *are* considering (5*∀)—as it happens, successfully—then you'll *have to* have some cognitive attitude towards it that wouldn't be available were Jack not to exist. You just won't be in a position to know a priori that you do.

BIBLIOGRAPHY

Bencivenga, Ermanno. (1986) "Free Logics" in D. Gabbay and F. Günthner, ed. *Handbook of Philosophical Logic*, Volume III. Reidel, Dordrecht, 373-426.

Burge, T. (1974) "Truth and Singular Terms" *Noûs* **8**, 309-25.

Lambert, K. (2001) "Free Logics" in L. Goble, ed. *The Blackwell Guide to Philosophical Logic*. Blackwell, Malden, Mass., 258-79.

Lehmann, S. (1994) "Strict Fregean Free Logic" *Journal of Philosophical Logic* **23**, 307-36.

Lehmann, S. (2002) "More Free Logic" in D. Gabbay and F. Günthner, ed. *Handbook of Philosophical Logic*, 2nd edition, Volume 5. Reidel, Dordrecht, 197-259.

Pryor, J. (2005) "There Is Immediate Justification," in Matthias Steup and Ernest Sosa, eds. *Contemporary Debates in Epistemology* (Blackwell, 2005).

Pryor, J. (2006a) "Hyper-Reliability and Apriority" *Proceedings of the Aristotelian Society* **106**, 327-44.

Smiley, T. (1960) "Sense without Denotation" *Analysis* **20**, 125-35.