

Horwich on Meaning

Allan Gibbard

University of Michigan

Ann Arbor, Michigan 48109

U. S. A.

Meaning comes eight years after *Truth*, Paul Horwich’s “minimalist” response to Pontius Pilate.¹ In the earlier book, the concept TRUE is characterized by a deflationist schema,

DOGS BARK is true if and only if dogs bark.

Small capitals indicate concepts, and so DOGS BARK is the proposition that dogs bark.² A closely related schema is Tarski’s:

‘Dogs bark’ is true if and only if dogs bark.

So put, however, the Tarski schema requires sticking to a single language—in this case, a single version of English.³ As for, say, the French statement ‘Les chiens aboient,’ its truth depends not only on whether dogs bark, but on whether those words mean DOGS BARK. The same indeed goes for the words ‘Dogs bark’ in someone else’s mouth, or in your own mouth yesterday; their truth depends in part on what they mean. Questions we might have wanted a theory of truth to answer, then, were passed on, in *Truth*, to a theory of meaning. “When I said ‘Dogs bark,’ was that true in that dogs bark? What was it about me and my surroundings that made those words true, apart from dogs’ barking?” Horwich’s minimalist theory of truth washed its hands of such questions, and passed the basin to a theory of meaning.⁴

¹ *Meaning* (1990); second edition (1998).

² Horwich uses capital letters for concepts; I choose small caps as easier on the eye.

³ This, of course, isn’t Tarski’s way of conceiving matters. Tarski argues that the object language can’t literally be the metalanguage, and requires that the metalanguage include translations of the object language.

⁴ Horwich is far from the only self-avowed “minimalist” on truth. Crispin Wright is another—but his account

Meaning, then, fills the gap left by *Truth*. Horwich's theory of meaning too, we learn, is "minimalist". This might initially lead us to fear that the theory consists of the deflationary schema,

'Dog' means DOG.

This schema needn't be useless; it does place a constraint on a theory of meaning: As said by you now in English, concerning the term 'dog' in your mouth now, the schema must hold—ambiguity and the like aside. This provides a constraint that a theory of meaning must satisfy. It doesn't, though, tell us what you are claiming when you interpret others, or yourself at another time or in counterfactual circumstances. If you are British, you can say this:

In American mouths, 'cotton' means not COTTON, but COTTON WOOL.

You will be right, but the deflationary schema wouldn't tell us why.

Happily, Horwich isn't "minimalist" in this way (11). So what is his theory? Quite simple: a Horwich book lays out its thesis in three pages (44–46)—and then spends its remaining centimeter of thickness fending off objections and exploring implications and relevant controversies. The sketch in *Meaning* is preceded by a fascinating disquisition on "pseudo-constraints on a theory of meaning". It isn't, though, followed by much elaboration or refinement; when it comes to sheer formulation of the theory, the sketch is pretty much what we get. (An "Obscurity" objection rates two additional pages, 57–59.)

The sketch expounds three theses: "(I) Meanings are concepts" (44). "(II) The overall use of each word stems from its possession of a basic acceptance property" (44). "(III) 'Two words express the same concept in virtue of having the same basic acceptance property" (46). To understand the theory, then, we need chiefly to grasp two notions: (i) *concepts*, and (ii) *basic acceptance properties*. I'll begin, though, supposing that we understand these, and look at how the theory operates.

'Schnee ist weiss,' in Hans's mouth, means SNOW IS WHITE—which is true. How could I settle what these German words mean? Roughly, I discover their basic acceptance properties, and I then discover that my own words 'Snow is white' share these properties. I know, then, that 'Schnee ist weiss' in Hans's mouth and 'Snow is white' in my own mean the same. Trivially, in my own mouth now, 'Snow is white' means SNOW IS WHITE.

differs so sharply from Horwich's that it would require a substantial treatment of its own. See Wright, *Truth and Objectivity* (1992).

In Horwich's system, this truism follows from his capitalization convention for naming concepts: the capitalized term 'SNOW', in my mouth now, refers to whatever concept 'snow', in my mouth now, means. Hence 'Schnee ist weiss' in Hans's mouth means SNOW IS WHITE; more idiomatically, it means that snow is white. And since snow is indeed white, what Hans says is true.

When Miyako tells her child "Inu wa hoe," we might roughly say she means that dogs bark. That, however, would be inexact. What she says has three words, whereas 'Dogs bark' has two. The meaning of a complex is built up from the meanings of its components, including operations of concatenation, inflection, and the like. I can't quite say in English what Miyako means, or at least not in terms of the capitalization convention. I could tell you of her meanings in a long-winded way, by stating the basic use properties of 'hoe' and her other words. This information, however, won't entail what to expect of dogs if Miyako is right. To do that, I'll have to learn Japanese, training myself up in exemplifying the basic acceptance properties that Japanese exemplify, sharing the dispositions which give their terms and elementary constructions the basic acceptance properties they have. Or more easily, I might train myself up in "Japalish", a language with English lexicon and some Japanese syntax. I'll then have such sentences at my command as 'Dog as-for bark' (where the 'as-for' applies to the preceding 'dog'), and I will be disposed to use such a sentence to get a child to expect dogs to bark. I can then say,

'Inu wa hoe,' in the mouth of Miyako, means DOG AS-FOR BARK.

Accepting this, I now treat Miyako as if she had said, in my new language, "Dog as-for bark." To see if she speaks truly, I ask myself, as I now put it, whether "dog as-for bark," and my answer is yes.

Horwich's theory, then, is deflationist, in that although it tells me how to evaluate an utterance for its meaning and for its truth or falsity, there's no telling from these directions alone what property constitutes an utterance's being true. We can't, Horwich tells us, "read off, for any given meaning-constituting property, which meaning it determines" (65). When Miyako tells her child "Inu wa hoe," I can't deduce solely from her propensities to usage and Horwich's theory what she is saying and what would make it true; for that, I must further train myself up and note my own propensities, using myself to model her. Some property does constitute meaning BARK, and Horwich's theory tells us how to find it. The property is whatever basic acceptance property governs my own use of 'bark'. (That is, it governs my use on an implicitly specified occasion. The capitalization convention that defines 'BARK' depends on this implicit specification, so as to rule out my use of 'bark')

on other occasions as a noun or with regard to trees and shins (81–85)—uses governed by different basic acceptance properties.)

As for meaning DOGS BARK, that consists in being composed as my ‘Dogs bark’ is composed, from elements with the same basic acceptance properties. To mean DOGS BARK is, I take it, to concatenate the plural of a term that means DOG with the third person plural of a verb that means BARK. Horwich’s account of how meanings compose, then, is trivial or nearly so, a point he stresses.

Where does that leave the semanticists’ stock in trade, “semantics” in the sense of following Tarski, determining how the reference and truth conditions of complex expressions compose?⁵ Semantics, we might think, must bear heavily on use; it bears, for instance, on when I will utter “Your dog is barking,” and when you will assent. A theory of meaning and use, though, as Horwich conceives it, won’t directly include semantics; it won’t include a theory of how truth and reference conditions compose. Semantic theorizing is a further activity. Again, then, Horwich follows a minimalist pattern, lopping off seeming responsibilities of the theory he is treating and passing them on. Use, he says, is to be explained in terms of basic acceptance properties, but to do this is not to engage in semantics. For to do semantics, one must have acquired the language or an exact translation; one must be properly trained up. To explain use in terms of basic acceptance properties, in contrast, if Horwich is right, one need speak only some language or other, a language that suits causal/explanatory needs.

Horwich’s scheme, we should note, gives me two ways to allude to the meaning of Miyako’s word ‘inu’. I can say, “‘Inu’ means DOG.” Or I can pick out its meaning by describing its basic acceptance property, the one that ‘inu’ and ‘dog’ share. ‘Inu’ means DOG, according to Horwich, just in case ‘inu’ and your term ‘dog’ share their basic acceptance property. Whether ‘inu’ means DOG, then, can’t be “read off” from its basic acceptance property; you can only know that it means DOG if your yourself have a term that means DOG, and know that it ‘inu’ share their basic acceptance property.⁶

For one’s own language of the present moment, one can do semantics with no thought of basic acceptance properties. For other cases, one does semantics by translating into a language of one’s own, doing semantics for one’s own language, and then transferring

⁵ I thank Jason Stanley for raising this question and for extensive, valuable discussion.

⁶ In Horwich’s vocabulary, the meaning and the basic acceptance property are distinct “properties”; the latter “constitutes” the former (46; see 2.2 and objection 4).

the results. And what constitutes a correct translation is a matter of basic acceptance properties.

Horwich's approach and execution display remarkable ingenuity, insight, and craft. He offers answers to questions that might have seemed impossible to handle within the kind of framework he develops—and all this with marvelously crisp, clear writing. His proposals may turn out to be immensely fruitful; I'll work in this essay to explore both the potentialities of the program and seeming obstacles to it.

The entire exposition I have been giving depends, it should be apparent, on the notion of a “basic acceptance property”. I don't think that Horwich tells us all we need to know about what such a property is. What he does tell us is this: Where w is a word and $A(x)$ is its basic acceptance property, “all uses of w stem from its possession of acceptance property $A(x)$ ” (45). The basic acceptance property “designates the circumstances in which certain specified sentences containing the word are accepted” (45). The property is explanatorily basic, accounting for all uses of the word (58). It is non-semantic and readily detectable (58); it can't be, say, that the word is used to refer to Napoleon, or that such-and-such brain process accounts for uses of the word (58). The meaning of a term, then, is a regularity; it explains causally. (Horwich thinks little of recent claims that “meaning is normative”, that the concept of meaning is a normative concept.)

Uses are, of course, only in part explained by meanings. Beliefs, desires—to inform or impress or flatter or protect secrets—lapses, glitches, limits on our capacities, and other factors enter as well into explaining why a person uses a word on an occasion. Horwich, though, ties meanings specifically to “basic acceptance properties” (or sometimes “basic use regularities”, 211). Meanings figure in explaining all aspects of use, but consist in one narrower aspect of use, acceptance of certain sentences in certain conditions.

How does this meet the worries of Quine, and of Kripke's Wittgenstein? Horwich responds to both; we might frame the debate as follows: On normal conceptions, all parties agree, uses of words and syntactic devices depend on multiple factors. Schematically, let these be (i) meanings, (ii) beliefs, and (iii) glitches. Glitches here will be errors, limits of capacity, and the like. This listing is by no means exhaustive, but these factors bear systematically on what sentences an individual i accepts: absent glitches, one accepts sentences with meanings one believes; I accept ‘Dogs bark’ if I believe DOGS BARK. Quine worries that even absent glitches, meanings can't be disentangled from what a person believes. (I might accept ‘Dogs bark’ because to me it means CATS MEOW and I believe CATS MEOW.) Kripke's Wittgenstein worries over glitches: What constitutes a glitch rather

than a refined aspect of meaning? (Might accepting ‘ $58 + 67 = 125$ ’ stem from a glitch, where ‘+’ means QUUS?) Where glitches clearly predominate, what constitutes meaning one thing rather than another? (With numbers too big to compute, what’s the difference between meaning PLUS and QUUS?)

Horwich has many things to say about Quine’s and Kripke’s arguments, but given his account of meaning, his central answer must come down to this: ‘Inu’ means DOG if its use is best so explained, if its use is best explained by the same acceptance property as best explains our own use of ‘dog’. The correct demarcations, then, between meanings, beliefs, and glitches are the ones that best explain use. This will allow some small amount of indeterminacy in what words mean, along with such things as a small set of words having meaning only with respect to each other. (An important and fascinating part of his response to Quine is to make distinctions among various purported phenomena in the vicinity of indeterminacy.) But it won’t, Horwich insists, allow the kinds of massive indeterminacy that worry Quine and Kripke.

The best explanation of Miyako’s pattern of acceptance and rejection of sentences, then, will impute meanings, beliefs, and glitches. Once we know what this explanation is, we’ll know what her words and sentences mean: they mean what the explanation says they mean. This seems eminently plausible, perhaps even truistic: words mean what they are best explained as meaning.⁷ To complete this account of meaning, then, we need two things: We need to understand being better or worse as an explanation—but this is a general problem in the theory of explanation, perhaps no special responsibility of a meaning theorist. Second, though, we need to make sure we understand what it is to explain a word as *meaning* something—to explain, say, ‘inu’ as meaning DOG. Explanations of acceptance, we are supposing, explain accepting a sentence as a matter of the meanings of the terms in it, what the person believes, and glitches. Beliefs, we might add, are explained in terms of such things as evidence, what others say and he takes their sentences to mean, and the like. What it would be good to have, then, is an account of what is special about the role of meanings in such explanations. What is the “meaning role” in an explanation of acceptance?

Marie accepts ‘Il fait chaud,’ and the explanation of this includes, glitches aside, not

⁷ Some who believe that “meaning is normative,” to be sure, will deny that meanings are defined as meeting purely causal-explanatory demands. I take up this slogan later, but if Horwich’s program works, there will be little reason to accept it; the chief arguments for it consist in discrediting the alternatives. Putnam, though, will point out that this characterization is itself normative, in that it adverts to talk of the best.

only that ‘Il fait chaud,’ means IT’S HOT, but that she believes that it’s hot—because she feels the heat of the day. Her meaning IT’S HOT consists in certain properties of its components, properties that play the meaning role in this explanation. What distinguishes this role? On this, Horwich is terse. A correct attribution of meaning to a word “offers the best explanation (when combined with other psychological laws) of the total use of the concept, including all the beliefs in which it appears” (147). But in this explanation, we must add, the property in question must play the “meaning role”, not a role that mixes meanings with beliefs or glitches—and we are asking after the difference in these roles. The meaning role is fundamental, we are told, but so, presumably, are certain principles of belief formation. Beyond that, the meaning property consists in an acceptance pattern, that certain sorts of sentences are accepted in certain conditions.

Horwich’s claim that meanings are determinate, then, amounts to this: For each word or syntactic device, there is a unique acceptance pattern with this quality: that for every use of the word, this pattern figures fundamentally in the best explanation of the use. Horwich’s actual claim is that meanings are determinate mostly; it is that this unhedged claim holds to a good approximation. This is a bold claim and a claim whose content is clear enough. It is largely empirical, a claim about explanations of language use. To get the rest of Horwich’s theory, just add: And this we call meaning.

Will this quell Quine’s qualms (as Horwich puts it, with his alluringly allusive alliteration)? “The whale is indeed a great fish!” exclaims Jonah to the people of Ninevah. He is committed, then, to thinking the whale a fish—and that’s wrong. But wait; we can’t judge this so quickly! Jonah isn’t speaking modern English, even if it is convenient to tell our story as if he were. He’s speaking Assyrian, presumably, but he also thinks in Hebrew; he accepts ‘Haaf dag’, which we are inclined to translate “The whale is a fish.”⁸ In Hebrew, after all, ‘haaf’ means WHALE, ‘dag’ means FISH, and (as all those italics in the King James translation teach traditional protestants) Hebrew needs no overt copula. Why, though, think ‘dag’ means FISH? Mightn’t it mean SWIMBEAST, in the sense of an animal streamlined for underwater swimming? (Understand this as a simple concept rather than as one formed by composition, in the way that MAMMAL is simple, though a mammal is defined as a warm-blooded animal that gives birth and suckles its young.) If that is what Jonah means, then he speaks truly: the whale is a swimbeast, and that is what he is saying. How do we set out, then, to decide what ‘dag’ means in Jonah’s mouth? That

⁸ Apologies: ‘Dag’ in Hebrew does mean FISH, but I need to learn what the Hebrew word for WHALE is, and revise accordingly.

is Quine's challenge. Horwich doesn't need to give us an answer, and he doesn't need to maintain that there is any determinate fact of this matter. We do, however, need to know what would be involved in getting an answer or finding it indeterminate.

Jonah applies 'dag' to paradigm fish, as I do the term 'fish'. Perhaps this is the basic acceptance property that governs our respective uses: seeing a fish, I accept 'That's a fish.' If so, then Jonah and I mean the same thing—and he falsely thinks the whale is a fish. He, though, applies 'dag' as well to whales, whereas to them I don't apply 'fish'. This is a difference in use, a difference in our patterns of acceptance; what explains it? Jonah applies 'dag' to apparent swimbeasts, and I might train myself up to do this with the term 'swimbeast'. For me, then, different acceptance patterns characterize 'fish' and 'swimbeast'; which of these fundamentally explains Jonah's uses? We differ in more than whether we lump whales with fish: I, after all, have been indoctrinated since childhood, "Whales are mammals, not fish," whereas Jonah has no term with any claim to mean MAMMAL. I know, and Jonah doesn't, that whales are warm-blooded, give birth and suckle their young, whereas paradigm fish are cold-blooded, lay eggs, and leave their young to fend for themselves. How, one of Quine's challenges runs, hold constant these differences in belief, and settle if Jonah's basic acceptance property for 'dag' is mine for some term, old or newly coined, in my vocabulary?

We might conclude that I, contaminated with modern scientific education, can't have a term with the same basic acceptance property, and can't acquire one without giving up science. In that case, I can't judge Jonah mistaken or not; the procedure for doing that requires me to get such a term by training up if I don't already have one. Alternatively, we might contend that, even if I don't already have such a term, I can acquire one. In that case, however, we need to know what constitutes success in doing so. Suppose I am trained up in some term w , a candidate translation for Jonah's 'dag'; the term w now has a use with me. I may then share some aspects of Jonah's pattern of acceptance for 'dag'. But I surely won't share his entire pattern. He and I just aren't well explained as sharing all our beliefs, and so there will be differences in what he accepts for 'dag' and I accept for w . Horwich tells us to look for whatever aspect A_j of his acceptance pattern best explains his entire use of 'dag', and likewise with me: A_g will best explain my use of w . Then we can ask if they are the same, if $A_j = A_g$. But is what most fundamentally explains Jonah's use of 'dag' what explains my use of 'fish', or 'swimbeast', or neither? We haven't so far learned how to answer this. Things would go far better if we could be seen as doppelgänger modulo our vocabulary, but we can't; our education and evidence

differ so greatly. This, I think, is one chief qualm of Quine's.

Horwich's critique fixes on a more extreme claim that Quine makes: that even with your doppelgänger, nothing settles that by 'fish' he means FISH. Horwich has fascinating things to say about this claim, things that I accept, but by answering the extreme claim, he thinks he is refuting widespread indeterminacy. This misses indeterminacy that arises because others just can't be interpreted as believing what we do. An adequate translation manual, Horwich says, is "a device of expectation replacement" (206); it "converts reliable patterns of expectation in the home context into reliable expectations abroad" 204e. "We are to have exactly the expectations we would have had if we were dealing with members of our own linguistic community—modulo the substitution of the corresponding foreign sentence for our own (205). Nice trick if we can accomplish it—but if their ways are more than superficially alien, no manual will accomplish this; any possible manual must leave us prone to greater surprises than at home. We can, of course, shape it to minimize surprise, more or less, but more than one manual may then qualify. One of Quine's challenges is to distinguish facts of meaning from facts of surprising belief. The Jonah example is a schematic case in point.

The problem remains if we push the case harder. More figures in the use of 'dag' or 'fish' than paradigm cases of recognition. Even people who, like Jonah, are untutored in Linnean taxonomy think in terms of folk-genus and folk-species: the raspberry is a species of bush, and the like. Folk taxonomy is essentialist, and to say all this is to allude to patterns of acceptance: a wolf in fish's clothing still doesn't count as a "dag". (So we translate Jonah as insisting, once we've ascribed meanings to everything but 'dag' in his vocabulary.) Such patterns of thinking emerge young; they are a cultural universal, and may be "wired in", provided for as adaptations in the human genetic plan for forming the brain.⁹ Jonah and other ancient Hebrews implicitly accept, then, that there is a hidden essence that characterizes prototypical fish. Linnean taxonomists—and, more recently and completely, cladistic taxonomists—have identified this essence. Whales don't partake of it.¹⁰

Jonah's word 'dag', we might then conclude, plausibly meant FISH, where a *fish* is an animal with the basic plan for reproducing of prototypical fish, evolved in a common

⁹ Atran (1990) finds cross-cultural patterns in classifying animals, and these might be human genetic adaptations for living in the world of our ancestors.

¹⁰ Unfortunately for Linneans, neither do sharks or lampreys. See Dupré (1993), 29–30.

ancestor and maintained, in many species, ever since. His acceptance of essentialist claims framed with ‘dag’ is part of what fundamentally explains the rest of his use. For these essentialist sentences are a part of his use, and they can’t be explained just by his recognition of paradigms.

Swimbeasts, though, also share a nature. They don’t specially resemble each other because of shared ancestry, but, in biologists’ terminology, they share major homologies; they are adapted to swimming by streamlining, muscular flat surfaces for propulsion, and the like. They have these features “by nature,” in that their respective genetic plans each provide for these features. This is a character they have essentially, in that, for instance, a wolf in fish’s or whales’ clothing doesn’t share it. The essentialist features of Jonah’s use of ‘dag’, then, I share both with ‘fish’ and with ‘swimbeast’.

Now, I could solve this quandary, perhaps, by learning ancient Hebrew myself, and then deciding, as I’d now be putting it to myself in my new, mixed language, “whether haaf dag”. Suppose, though, I say not. We can now ask whether it is really Jonah’s Hebrew I have learned, or a variant in which ‘dag’ means FISH instead of SWIMBEAST. After all, by learning Jonah’s language, I haven’t come to match his beliefs; mere language learning doesn’t purge me of scientific knowledge. We still need help from Horwich in figuring out whether, indeed, I have learned Jonah’s language or some other.

Horwich addresses aspects of this worry in his treatment of implicit definitions (Chap. 6). Phlogiston theory consists in a set of axioms, and the meaning of the term ‘phlogiston’ that figures in these axioms must help explain theorists’ acceptance of the axioms. One proposal for the term’s meaning is this: the basic acceptance regularity of the term consists in users’ accepting these axioms—call them $T(\text{phlogiston})$. Those who reject the theory, though, also understand it; they too can use the term ‘phlogiston’ with meaning. Now, accepting or rejecting the theory consists in accepting or rejecting its Ramsey sentence $\exists x T(x)$. And even those who reject the theory still accept a hedged conditional, $\exists x T(x) \rightarrow T(\text{phlogiston})$. This acceptance regularity opponents share with adherents.

Perhaps this, then, is the basic acceptance regularity for ‘phlogiston’, for adherents and opponents alike. They’d better all share the same basic acceptance regularity, after all, or else they mean different things by the term. They can’t then *engage* using the term, in that they can’t use the term directly to disagree: they won’t be disagreeing if, say, one says “Wood contains phlogiston” and the other says “Wood does not contain phlogiston.” What, though, if two adherents accept phlogiston theory in different versions: Joe accepts

T_j (phlogiston), whereas Flo accepts T_f (phlogiston). Can these two engage using the term? Perhaps the term then gets its meaning from a core theory T^* (phlogiston) that Joe and Flo share, of which T_j (phlogiston) and T_f (phlogiston) are elaborations. They both accept the conditional, $\exists x T^*(x) \rightarrow T^*$ (phlogiston), and so can those who reject phlogiston theory in any form. Any two, then, will mean the same thing by the term, and can using it.

It is hard to see, though, how to derive this result from Horwich's theory. Why is the basic acceptance regularity that explains Joe's use the one we need, his acceptance of the conditional $\exists x T^*(x) \rightarrow T^*$ (phlogiston) with the minimal core theory? He also accepts $\exists x T_j(x) \rightarrow T_j$ (phlogiston); mightn't this explain his individual use better? Indeed he accepts the unhedged claim T_j (phlogiston); isn't this what most fundamentally explains Joe's use of the term? If so, then correspondingly with Flo—and the two, in holding different theories of phlogiston, mean different things by the term. They then can't engage using the term: no two theorists can straightforwardly debate the nature of phlogiston.

This, of course, isn't a problem peculiar to Horwich's own theory. It is in effect Quine's problem in one of its versions, the familiar way that we seem both forced to holism in a theory of meaning and unable to live with the consequences. Perhaps these phenomena vitiate all attempts to make sense of thinkers' agreeing and disagreeing with each other. Horwich does escape the most rampant of holism: not everything that Joe accepts counts as basic—"There's phlogiston in that log" doesn't, for instance, because it plays no fundamental role in explaining his whole pattern of acceptance. We do, however, get the result that those who disagree in their basic theories can't engage using a theoretical term like 'phlogiston': neither can deny what the other says.

Why, though, believe that Joe and Flo do engage when they use the term 'phlogiston'? Because they converse, and in conversing they treat each other as engaging, as meaning the same things by their terms. In this they might of course be mistaken—as perhaps were theorists of *forcia*, when some meant energy by the term and others meant momentum.¹¹ But the social phenomenon of purported engagement does seem here to carry weight in the best interpretation of Joe's language. It carries weight especially if it continues when Joe and Flo realize the full extent of their differences. It isn't, then, just an aspect of what Joe accepts that explains what he means; it's how he engages with Flo, understanding her—as he might put it—to mean PHLOGISTON by 'phlogiston'. It's his evincing this interpretation of Flo's words that moves us to treat them as sharing a meaning for the

¹¹ Debates on "force".

term ‘phlogiston’, based perhaps in a common core theory $T^*(\text{phlogiston})$. And with Antoine the phlogiston sceptic, it’s their mutual interpretation as meaning PHLOGISTON by ‘phlogiston’ that brings us to the hedged form $\exists x T^*(x) \rightarrow T^*(\text{phlogiston})$ as explaining their common use. Patterns of acceptance, then, aren’t all that explains use; patterns of apparent engagement do so too.¹²

Still, suppose a group of phlogiston theorists converse only with each other. Then there’s nothing about their conversational interactions that need make us think that they share their concept PHLOGISTON with phlogiston skeptics, that the regularity that basically explains their use is accepting the hedged $\exists x T^*(x) \rightarrow T^*(\text{phlogiston})$ rather than the unhedged $T^*(\text{phlogiston})$. That will mean, on Horwich’s account, that we who exemplify no corresponding acceptance regularity can’t speak their language, and so can’t deny anything they say using the term ‘phlogiston’. When they say “There’s phlogiston in that log,” we can’t reply “No, there isn’t;” we’ll have to say something more roundabout. I’m not at all sure that this is a damning consequence; certainly I don’t have an alternative account that lets me avoid it. What recommends taking as their basic use regularity the hedged $\exists x T^*(x) \rightarrow T^*(\text{phlogiston})$, though, isn’t some matter of how to explain their use. It is that this attribution lets us regard the widest possible group as able to engage with them using the term ‘phlogiston’. In particular, this attribution lets us regard us ourselves as able to engage with them. This is a kind of charity: not so much attributing truths to them, but at least attributing to them thoughts couched in terms we can share.

What, then, of claims that “meaning is normative”? Meanings of course underlie certain oughts, as Horwich points out, just as any kind of fact can underlie an ought: rain underlies my finding that I ought to take an umbrella. Being normative in this loose sense wouldn’t threaten a naturalistic account of meaning like Horwich’s; on this point Horwich is clearly right. What kind of claim of normativity, then, might go against Horwich’s naturalism? One would be that “means implies ought”, that meaning claims are tied analytically to normative claims, in a way that can’t be explained if a naturalistic account of what ‘means’ means is correct. Why think such a thing? Suppose we can’t meet Quine’s challenge by saying, in naturalistic terms, what ‘means’ means. What’s at issue, we might then try asking, when people disagree on what a word means? Perhaps the issue is not

¹² This moves us in the direction of Brandom (1994), who finds meaning in the norms that people apply to each other. It doesn’t by itself, though, require us to say that Joe’s meanings are social rather than “in his head”. Something in his head, after all, grounds his dispositions to hear Flo and Antoine as meaning PHLOGISTON by ‘phlogiston’.

how the word is actually used and what explains this use. Perhaps it is instead how *to use* the word—and this in effect is an *ought* question, a normative question of how to use it. Perhaps it's a question of what to accept and what to reject, and not a question of what people do in fact accept and why. Horwich recognizes only pragmatic questions in this vein, questions of advantage and disadvantage. Suppose, though, I ask myself “what to conclude” from a body of evidence—say, the evidence on human origins. This wouldn't mean, on its most natural reading, what to brainwash myself into concluding in order to win election to a bible-belt school board. That, we might say, is a matter of what to *want* to conclude in light of my aims. There do seem to be normative questions that are not directly pragmatic questions, not questions of what to want and what to pursue. When Flo says “That log has phlogiston,” I can ask whether to accept what she says, and the pragmatic advantages of going along with a straight face don't bear on the question. What she means does bear: if she means that the log has phlogiston and I know quite well that it doesn't, her claim is one to reject. If she means that the log can burn, her claim is one to accept.

Horwich, as I say, finds no sense in non-pragmatic norms, and no gap in a causal/explanatory role account that needs filling by a claim that the concept of meaning is a normative concept. An interesting claim that the concept of meaning is normative would require two things: finding that there indeed is something at issue with meanings beyond how to explain use, and showing that the issue can somehow be explained as one of oughts. Whether such a program is promising isn't to be settled quickly.

Horwich's theory takes the form that has become known in the metaethics literature as expressivist. Consider an expressivist for ethics: To explain the meaning of the concept GOOD, he doesn't attempt to offer an analytic definition. Instead, he explains what it is to *think* something good. To think something good, he might say, is to approve of the thing. This expressivist, then, explains the concept GOOD psychologically, by starting with the psychic state of deploying the concept. Most expressivists have thought that this method of explanation suits some concepts and not others; indeed expressivism for a realm of concepts is often classified as a form of irrealism for that realm. Horwich applies the expressivist's stratagem to every concept whatsoever—in a specially ingenious and complex way. To explain the concept DOGS BARK, we explain what it is to *think* DOGS BARK. This consists in combining *DOG*, pluralization, BARK, present third personalization, and predication. Deploying each of these conceptual elements consists in satisfying a basic

use property. Horwich's scheme for explaining a concept, then, is to offer a complex account of what it is to have thoughts that employ the concept and related ones.

Expressivism has of course been widely debated, and we can ask how lessons of these debates bear on Horwich's own expressivist program. At the root of meaning, Horwich places acceptance: we explain the meaning of a term by conditions under which certain utterances that include the term are accepted. Not all uses of language amount to accepting anything. "Hiyo! Bob" can be used to hail Bob, and we could even imagine a language that puts this syntactically in the form of predication: I hail Bob by saying, "Bob is hiyo."¹³ Horwich's account of meaning, once filled out, would have to explain to us what acceptance consists in. One accepts strings of words, we might say; one accepts, say, 'Dogs bark.'. (More idiomatically, one accepts what those words say in one's language, but that can't be the starting point for an account of meaning, since "what those words say in one's language" is just what one's words mean, and that is what is to be explained.) Which uses of words constitute accepting a string of words, and which—like saying "Bob is hiyo"—don't? What is acceptance. I don't know if there is an easy answer, but it would help us to understand Horwich's account if we had the answer.

Accepting a string of words constitutes believing what those words mean. A crucial feature of belief is that different people, at different times, can believe or disbelieve the same thing. If Jill believes DOGS BARK, others can agree or disagree with her, and she herself can later change her mind. Not all uses of language have this feature; one can't, for instance, agree or disagree with "Hiyo! Bob"—and if "Bob is hiyo" were part of the language, one couldn't agree or disagree with that. "Yowee!" can express a headache, and in a modified language, we might substitute "I am yowee" to express a headache. These words might express a headache, as "Dogs bark" expresses another state of mind: the belief that dogs bark. But whereas one can disagree with a belief, one can't disagree with a headache. What is the difference? The possibility of agreement and disagreement seems crucial to belief, and Horwich's account raises the puzzle, how could anyone agree or disagree with the kinds of states that Horwich describes.¹⁴

¹³ The point and example are from Dreier (1996), who discusses Horwich (1994) on expressivism.

¹⁴ This may be related to a critique of expressivism by Nicholas Unwin. Unwin argues that an expressivist like me cannot account for negation. An atheist denies that any gods exist; an agnostic permanently rules out accepting that gods exist, without denying that any do. An expressivist for value and the like, Unwin argues, can't account for such a distinction; she can't say what's different between denying that pleasure from heroin is good and just permanently ruling out accepting that it is good. Denying a claim consists in disagreeing with belief in it, and so Unwin's challenge amounts to asking whether expressivists can

We could put the puzzle this way: People can agree or disagree with my belief that dogs bark; that's truistic. Horwich has an account of what having this belief consists in; it consists in a complex pattern of acceptance properties. Suppose, then, we just consider this pattern as theorists. We tell Martian scientists about this pattern and related patterns, and discuss these patterns with them. Can we explain to the Martian what it is for others to agree or disagree with Jill when Jill exemplifies this pattern? The emotivist C.L. Stevenson, to show that moral beliefs might consist in attitudes, started by arguing that there can be disagreement in attitude. Attitudes, then, can be treated by others as true or false, in effect; I treat your disapproval as true if I share in your disapproval. Headaches aren't like this: if we both have headaches, we don't thereby agree about something. Getting an expressivist program off the ground involves showing that a set of states of mind—attitudes, for instance—that can be explained independently of attributing content to them mimics, in crucial ways, straightforward beliefs. Crucially, that includes that people can disagree in those states of mind—people can disagree in attitude as well as in belief. Horwich, in effect, is claiming that a certain kind of state of mind, identifiable independently of saying that it is a belief in DOGS BARK, turns out to be a belief in DOGS BARK. To show that it fits the job requirements, must he show that people, being in such a state, can thereby be in agreement or disagreement?

I'm not sure this is an objection; I mean it more as a request for guidance—made, of course, with some worry that the guidance might impossible to find. I don't know whether Horwich really owes us a naturalistic account of disagreement in belief. His account has two quite different parts. On the one hand, there's the part we could share with Martians who are capable only of thinking in naturalistic terms. We can tell these Martians all about patterns of basic acceptance properties (provided, still, that we can explain to them what acceptance consists in). The other part of the account I get by matching other people's basic acceptance properties with my own; only then can I judge that by 'Dogs bark,' Jill means DOGS BARK. As I have said, it is only at this second stage that I can myself agree or disagree with Jill. I can now judge that what she thinks is true, whereas my Martian colleagues can't.

This suggests an answer to the puzzle. It is only at the second stage that I can judge that Jack and Jill agree or disagree. They disagree if Jill believes DOGS BARK and Jack believes DOGS DON'T BARK. More generally, they agree if they believe the same thing, and disagree if one believes the negation of what the other believes. I identify a syntactic

coherently include disagreement in their picture.

device as negation by finding it in myself. (Something further must be said: if Jack believes I'M HOT and Jill believes I'M NOT HOT, they don't thereby disagree; Jack believes he's hot and Jill believes she isn't. Accounting for this ubiquitous aspect of agreement and disagreement is important, but whether Horwich should face any difficulty in principle in doing so I won't try to investigate.)

[xxx how conclude this section.]

Horwich's program, if it succeeds, tells us what meaning DOG consists in, and likewise for each other concept. Even if this succeeds, this might not tell us what 'means DOG' means. Good, if hedonists are right, consists in pleasure, but 'good', as G.E. Moore argued, doesn't mean *PLEASANT*. For a perfectionist can agree that heroin is pleasant but disagree that it is good. Even the hedonist can distinguish which of her views this hedonist accepts and which he rejects. Can Horwich tell us not only what meaning DOG consists in, but what 'means DOG' means? What is at issue when people disagree about meanings?

Presumably, an account of what 'means' means should take the form, with Horwich, of a basic acceptance property for 'means'. Or perhaps it should consist in a basic acceptance property for 'means DOG', another for 'means GOOD', and so forth. It would be interesting to discover a plausible candidate for such a basic acceptance property.

In any case, Horwich isn't claiming that his theory of meaning offers a synonym for means. He's telling us how to identify such things as the property that constitutes meaning DOG by 'dog'. If someone offers an alternative account, is something clear at issue between this theorist and Horwich?

Horwich speaks of ascribing meanings as a device of "expectation replacement": knowing that with Tijean, 'chien' means DOG, I can treat him, when he says 'chien', as if he were a family member saying 'dog'. Horwich's his account of meaning is tailored to serve this function. We ascribe meanings, though, not only to form expectations. For one thing, we have those reactions that constitute agreeing or disagreeing. If Tijean says "Tiens, une biche" and gestures toward a doe, I object if I think he means "Hey, a dog," but not if I think he means "Hey, a doe." If ascribing meanings consists in reactions that go beyond sheer expectations, the meaning of meaning might lie in these reactions. Horwich's approach would allow this, if the basic acceptance property of 'means' involves more than expectations. (I should say, though, that my own attempts to work out what 'means' means along expressivist lines have not been successful. Still, if Horwich is right,

then *some* expressivist account of what ‘means’ means must be correct, if, as I claim, Horwich thinks that all meanings are to be elucidated expressivistically.)

In short, I remain puzzled by aspects of Horwich’s account. But I haven’t identified a crucial obstacle to the account’s working. Still, there’s a big question left even if Horwich is entirely successful in the task he sets himself. We still don’t know what ‘means’ means. If Horwich is right, this should have an answer. One kind of answer is expressivistic; it consists in identifying a basic acceptance property, either for ‘means’ or for such units as ‘means DOG’. Leaving open this question means that space is left, in Horwich’s book, for claims that the concept MEANS is a normative concept, or that in some other way, meaning attributions consist in more than a pattern of expectations. Whether such claims are right, nothing that Horwich says—or that I can find to say—settles.

References

- Atran, Scott (1990). “Cognitive Foundations of Natural History: Towards an Anthropology of Science” (Cambridge: Cambridge University Press).
- Brandom, Robert (1994). *Making It Explicit* (Cambridge, MA: Harvard University Press).
- Dreier, James (1996). “Expressivist Embeddings and Minimalist Truth”. *Philosophical Studies* **83**, 29–51.
- Dupré, John (1993). *The Disorder of Things: Metaphysical Foundations of the Disunity of Science* (Cambridge, MA: Harvard University Press).
- Horwich, Paul (1990). *Truth* (Oxford: Oxford University Press). Second edition, 1998.
- Horwich, Paul (1994). “The Essence of Expressivism”. *Analysis* **54**, 19–20.
- Horwich, Paul (1998). *Meaning* (Oxford: Clarendon Press).
- Unwin, Nicholas (forthcoming). “Norms and Negation: A Problem for Gibbard’s Logic”. *Philosophical Quarterly*.
- Wright, Crispin (1992). *Truth and Objectivity* (Cambridge, Mass.: Harvard University Press).