Horwich on Meaning


Allan Gibbard  
*University of Michigan*  
*Ann Arbor, Michigan 48109  
U. S. A.*

[Preliminary draft. Please do not quote or cite.¹]

In 1990 Paul Horwich published an intriguing but frustrating book *Truth*.² He stated his entire theory in few short pages, and then spent much of the rest of the book responding to objection after objection. He labeled his briefly described theory “deflationary”.

The deflationary schema for truth is just this: to say “It’s true that dogs bark” just amounts to saying that dogs bark. A fully deflationary theory of truth would presumably maintain that there’s no more to the concept of truth than this. Deflationary theories are of course nothing new, but it has been widely thought they couldn’t be made to work. A fully deflationary theory fails to account for generalizations, such as “Nothing de Gaulle said that day was true.” It doesn’t account for the truth in languages other than one’s own, as in “De Gaulle’s claim ‘Les chiens aboient’ was true.” It doesn’t account for truth in other contexts, as in “Churchill’s claim ‘I’m tired’ was true.” Horwich’s own theory, though, addressed these problems. Truth is explained by an axiom schema, and the axioms stand in logical relations to generalizations and the like. De Gaulle’s words were true if they meant something true, such as that dogs bark. Churchill’s claim was true if it meant something true, such as that he was tired. Armed with this theory, Horwich denied that truth is a substantial property, and offered deflationary explanations of claims that might seem to require a substantial property of truth. The “value of truth”, for example, amounts to such things as the value of believing that dogs bark just in case dog do bark—and similarly for “aiming at truth”.

The effect of all this was to load the burden of explaining truth onto a theory of meaning. “Les chiens aboient” is true in that (i) it means that dogs bark, and (ii) dogs

¹ This draft is much too long, and some notes and references remain to be filled in.
bark. The fully deflationary part of the theory is, in effect, this: if something means that
dogs bark, then it is true just in case dogs bark. What remains is to explain what it is
to *mean* that dogs bark. What made de Gaulle’s words “Les chiens aboient” mean that
dogs bark? What makes correct the translation of de Gaulle’s “Les chiens aboient” into
my “Dogs bark”? That’s a matter for a theory of meaning, and the book *Truth* left the
question untreated.

Then came the book *Meaning* (1998), with a theory of meaning likewise termed
“deflationary”. Now on analogy with truth, a *fully* deflationary theory of meaning, we
might think, would take this form: “Dogs bark” means that dogs bark. That is a crucial
part of Horwich’s account, and he has a special notation for it. The word ‘Dog’ means
DOG, where expressions in small caps denote concepts. Still, this fully deflationary part
of the account won’t tell us that de Gaulle’s word ‘chien’ means DOG. Horwich fills the
gap with a theory of correct translation, a theory of synonymy.

He calls his account a “use theory” of meaning, and works to vindicate Wittgenstein’s
dictum that the meaning of a word is its use. As he puts it in his most recent book
*Reflections of Meaning* (2005), the theory goes as follows: For a word, the meaning “is the
common factor in the explanations of its numerous occurrences” (26). This, he argues, is
“the (idealized) law governing its usage.” And what will such a law look like? The
law, Horwich answers, “dictates the ‘acceptance conditions’ of certain specified sentences
containing it” (26). It thus gives a *basic acceptance property* of the word. For the word
‘dog’, for instance, the basic acceptance property might roughly be that ‘That’s a dog’ is
accepted with attention focused on a nearby dog.

For ‘basic acceptance property’ I’ll write ‘BAP’, so that in the case of a word,
according to Horwich, its meaning is its BAP. Two words thus mean the same if they
have the same BAP. Or rather, for them to mean the same, their acceptance properties
must be analogous: De Gaulle’s word ‘Chien’ means DOG in that its BAP parallels that
of my word ‘dog’. That leaves the meaning of complexes like ‘Dogs bark,’ but for this,
Horwich is deflationary indeed. Its meaning is just the structure of the word meanings
that compose it. The compositionality of meaning puts no severe constraint on a theory
of how expressions get their meanings (10).

I myself, I should note, have a special interest in theories that characterize meaning in
Horwich’s way. As Horwich has noted, my own theory of the meanings of normative terms

---

3 In *Meaning* he uses full-sized capitals, which I find obtrusive, whereas the small caps convention in
*Reflections on Meaning* I find pleasant.

4 Page references in parentheses are to *Reflections on Meaning* (hereafter, “*Reflections*”), unless otherwise
indicated.
like ‘ought’ might be put in Horwich’s form.\(^5\) I call my theory *expressivist*: instead of offering a straight definition of a term like ‘ought’, the expressivist elucidates its meaning by saying what the psychic state of accepting an ought claim consists in. Older expressivist for morality like Ayer not only adopted this strategy for elucidating meanings but also denied that ethical claims are propositions, but current “expressivists” drop this negative claim, and so are expressivists in a weaker sense. Horwich is, we could say, an expressivist in this weaker sense for every meaning. And so is anyone else with a functional role or conceptual role theory of meanings.\(^6\)

What’s new in Horwich is not that his is a use theory. He puts such a theory in an especially clear and perspicuous form, but as he points out, it amounts to a conceptual role theory of meaning, and conceptual role theories have been around for some time. The particular form his use theory takes is indeed his own, but what I find most striking is his campaign to show that features of meaning that have been thought profound, and have been thought to contrain severely the form that an adequate theory of meaning could take, can be deflated away. Here are some of his deflating contentions: The “value of truth” consists simply in the value of such things as believing that gold glitters just in case gold glitters. The “aim of truth” consists in such aims as to accept that gold glitters just in case it does. The “problem of error” raised by Brandom, Kriple, and others is a pseudoproblem. There’s no reason to think that the concept of meaning is a normative concept. The compositionality of meanings requires no deep explanation. Interesting claims can’t be vindicated as true simply in virtue of their meanings—nor, indeed, can such uninteresting claims as that bachelors are unmarried. These contentions are trenchant and controversial. They don’t all depend on the particular form that Horwich’s theory of meaning takes, but some of them Horwich may be almost alone in maintaining.

His latest book, *Reflections on Meaning* (2005), combines new material with material already published, with light or heavy revisions. It starts out with a quick and perspicacious handbook guide to “The Space of Issues and Options” and a restatement of his “use” theory of meaning. Then in the remaining chapters, he tackles error, vagueness, norms of meaning and truth, semantic vindications, the language faculty, and compositionality. The result is a highly illuminating presentation of the theory and some of its implications.

---


\(^6\) Jason Stanley pointed out to me some years ago that what I was calling “expressivism” amounted to a functional role theory of the meanings of normative terms. Paul Boghossian convinced me to revise an earlier formulation of this point. For my own version of expressivism and my treatment of whether normative claims are propositions, see my *Thinking How to Live* (2003). Blackburn’s *Ruling Passions* is another treatment of expressivism that is not committed to denying that there are ethical propositions.
The new book convinces me on some points where I was doubtful before. In this essay I’ll chiefly be discussing this book, but I’ll allude back as well to the earlier books *Truth* and *Meaning*. I find much of what Horwich says persuasive and all of it helpful and thought-provoking. I’ll focus, of course, on the points where I’m puzzled or where, at least tentatively, I disagree.

1. Truth and Reference

Truth conditions and reference, Horwich insists, can’t be “read off” from meanings. In one sense they can, to be sure: the word ‘dog’ refers to dogs, and the sentence ‘Dogs bark’ is true in that dogs bark. This is the fully deflationary part of his account of truth and reference. In another sense, though, reference and truth-value can’t be read off from meaning, if Horwich is right. This is the feature of his theory that Horwich himself labels “deflationary” (19). The concept *DOG* is the BAP of the word ‘dog’, namely the property that putting the word in a context meaning *THAT IS A* . . . makes a sentence that is accepted with attention focused on a nearby dog. The meaning and this property are identical, and so we have two ways of designating one and the same meaning: the small caps designation ‘dog’ and an empirical description of the property (roughly, “the property that ‘That is a dog’ is accepted with attention focused on a dog.”) The two are coreferential but not synonymous, as with the terms ‘water’ and ‘H₂O’. Just as chemical structure of water can be read off the term ‘H₂O’ but can’t be read off the term ‘water’, though both terms name the same substance, so the reference of the word ‘dog’ can be read off the term ‘DOG’ but not off the codesignative description of the very same property.

The same with truth conditions: we can specify the meaning of a sentence by describing the BAPs of its components and saying how it is built up from those components. From this, however, even together with the ways the world is, there’s no general way to derive whether the sentence is true. If, then, de Gaulle once said, “Les chiens aboient,” how can I assess his statement as true or false? I don’t look to Horwich’s theory, cite the fact that dogs bark, and derive a verdict. Rather, I first note that his sentence meant *DOGS BARK*. In other words, his sentence meant the same thing as my own sentence ‘Dogs bark’ means, in that the two are constructed in a parallel way from words with parallel BAPs. Next, I note that indeed dogs bark, so that *DOGS BARK* is true. Thus de Gaulle’s words were true. What I have done is not apply some general condition I have formulated as to when exotic sentences are true and when they’re not. Rather, I

---

7 While I’m praising these books, I should mention by way of disclosure that Paul Horwich is a good friend. I was impressed by his book *Truth*, though, before I knew him at all well, and in his own honest spirit, I’ll try not to mute any criticisms or reservations I might have.
use the account of what thought de Gaulle was voicing to check that I’m having the same thought, and then check whether that’s the way things are. I don’t apply a theory of truth conditions; I simulate and then assess. I simulate his thinking by thinking in words that mean the same as his meant.

According to Horwich, then, there is no specifiable non-semantic relation that holds in general between a word and its reference, or between a sentence and its truth conditions. There are, though, specifiable particular meaning properties such as meaning DOG and meaning DOGS BARK. And there is a specifiable relation of synonymy, of meaning the same thing: it consists in being composed in the same way from words with parallel basic acceptance conditions. That reference and truth conditions can’t, in this sense, be read off meanings is a feature that Horwich’s theory shares with others such as Quine’s. Quine too addressed the nature of meaning by attempting a theory of synonymy—though, as he saw it, with success that could only be limited. Any theory of meaning whatsoever that includes a theory of synonymy would allow for assessing truth in Horwich’s way. Think something synonymous yourself, and then assess its truth in the fully deflationary way.

Some theories of meaning offer more than synonymy. Even Horwich’s own theory does: it tells us what sorts of things meanings are; it designates certain entities as meanings—namely, BAPs and structures comprised of them. Any theory of synonymy could of course be extended trivially to declare certain objects meanings, so long as synonymy is explained as an equivalence relation. We can follow Quine, and say that meanings are equivalence classes under the relation of synonymy. Horwich’s objects, though, are more theoretically interesting than mere synonymy classes. Where Horwich balks is with meanings as objects that, in some canonical way, settle truth conditions or reference.\(^8\) Even, though, for a theory that goes further in its ambitions, not only specifying as meanings properties that are open to empirical investigation but also saying how these properties determine truth and reference, the fully deflationary part of Horwich’s theory will provide a test. I test the ambitious theory by applying it to my own brain, dispositions, and surroundings right now, seeing what, according to it, is the property that constitutes what I mean by ‘Dogs bark’ and what truth conditions the theory accords that property. The test is whether those conditions are that dogs bark—and likewise for other sentences.\(^9\) In that sense, a theory of synonymy joined with the deflationary test constitutes a minimum for a theory.

---

\(^8\) I take it that like Frege, we can regard truth value as a special case of reference. Strictly, then, I should speak of referents and of the special case where the referent of an expression is a truth value, or I should speak of reference conditions and their special case, truth conditions. Being sloppy in this regard, though, seems to me to help readability.

\(^9\) Kripke in *Wittgenstein* (1982), speaking for Wittgenstein, claims that any dispositional account of meanings fails this test.
of meaning. Any non-debunking theory of meaning will entail what synonymy consists in, and to be successful, the theory must pass the deflationary test.

This way of assessing truth works only when I’m capable of thinking something synonymous. Sometimes I’m not. The reason may be trivial: Some languages don’t have articles like ‘a’ and ‘the’ or such grammatical devices as pluralization. Japanese ‘Inu wa hoe,’ according to Horwich’s theory, doesn’t really mean DOGS BARK, because its elements and compositional structure are different. To assess it for truth, I’ll need to learn a language with the same compositional structure. I might learn Japanese, so that Japanese words in my mouth acquire the same BAPs they have in Japanese mouths. I might learn, for my own purposes, a kind of broken English with the BAPs and compositional structure of Japanese but with English lexicon, as with “Dog as-for bark.” What if you learn this language and disagree with me in it, though we agree on the facts as couched in English? That will presumably be of a piece with disagreeing in our own language, and Horwich’s theory of synonymy won’t tell us who is right or what the dispute hinges on.

In some cases, though, a language will be too abstruse for me to learn—as with the language of advanced physics. How can I then assess its claims for truth? Not in the direct way, since I’m incapable simulating those who speak the language. The things we rightly accept, though, Horwich will say, go far beyond what the BAPs of our words settle by themselves. I can look at fresh-looking hoofprints on a path and think that a horse went by today—and this thought may well be justified, even if what’s meaning-determining for my word ‘horse’ is sentences I accept only with a horse right there in front of me. This point extends to TRUE: once we have the concept, we can justifiably believe things true even if we can’t ourselves, on this occasion, apply the deflationary schema that gives ‘true’ its meaning.

All this, as I say, is what follows for truth if what we have to work with is synonymy plus deflation. Don’t we need more, though? Truth plays a systematic explanatory role, we might object, to which this doesn’t do justice. Good science “tracks” the truth in a way that pseudo-science doesn’t; don’t we need a more substantial property of truth to explain this? Philip Kitcher has raised this problem with me, or something like it. Horwich, though, I think, can answer this. In the first place, some BAPs taken individually tend to truth. When I think “That’s a dog” with my attention focused on a nearby dog, I exemplify the BAP that makes my term ‘dog’ mean DOG, and I thereby think a truth. Other BAPs explain the preservation of truth in our inferences. When I start with true beliefs and apply modus ponens, I exemplify the BAP that gives the conditional ‘if . . . then’ its meaning, and draw a true conclusion. Not all BAPs, though, yield truths so automatically. I call “good” those things I favor, thereby exemplifying the

---

10 Philip Kitcher has raised this problem with me, or something like it.
BAP for good, but this is truth-conducive only because I tend to favor what’s good. I infer a nearby dog from hearing its bark, and think truly, but BAPs alone don’t explain my inference. I rely on induction, and to explain my reliability we must cite the reliability of induction. In general, then, to explain our tracking truth, we must look to each of our epistemic tendencies individually, whether founded directly in BAPs or not, and tell its own story.

Still, why the general tendency to truth in our thinking, at least in many realms? Where there’s an explanation, Horwich convinces me, deflationary truth allows it. First, of course, deflation explains why we will regard the things we think true to be true. Deflation and learning from experience can explain why, ideally, we will tend to change ways of thinking that lead to conclusions we find to be false. Why, though, will we mostly be right? For not all our epistemic tendencies will there be a substantial answer to this question. Why do we favor the good? Why do we expect the expectable, and why do the things we expect often turn out to happen? In the history of philosophy, vindications of these epistemic tendencies have proved hard to find—unless they are trivially self-congratulatory, as in “We favor happiness, and behold, happiness is good,” or “We expect the sun to rise tomorrow, and wait and see, it will.” It is part of Horwich’s account of the meaning of ‘true’ that the term can be used to express generalizations, and these will include whatever generalizations explain our tendencies to get things right. As for these generalizations themselves, they will have a variety of grounds. In some cases, they will have no further ground at all.

I have another misgiving, though, about Horwich’s account of meaning, a misgiving of quite a different sort. The meaning of a word is its BAP, and this consists of being ideally prone to accept certain sentences in certain conditions—and perhaps to reject others. Suppose, then, my friend Nancy speaks English but with an extra word ‘nath’. The properties that explain her deployment of this word are that she is (i) prone to accept ‘Tuesdays are nath,’ and (ii) prone to reject ‘Wednesdays are nath’. She professes not to know which other days are “nath” and which are not. Perhaps, she tells me, that isn’t a thing we can ever know. Take, now, a sentence in her idiolect, ‘Sundays are nath.’ What does its truth or falsity consist in? To assess its truth directly, I have to learn this piece of her idiolect myself, acquiring a new word—it might as well be ‘nath’—with this BAP. Doing this isn’t a matter of forming new beliefs in my old vocabulary, but in somehow training myself to react with a sense of automatic understanding to sentences with ‘nath’, and to accept effortlessly ‘Tuesdays are nath’ as I think it and reject ‘Wednesdays are nath’. Having a word and accepting or rejecting a sentence underived are familiar psychic states that Horwich’s theory takes as a starting point.\footnote{He offers a brief account of acceptance in \textit{Meaning}, pp. 94–96.} I now say, helping myself to

\footnote{He offers a brief account of acceptance in \textit{Meaning}, pp. 94–96.}
deflation, “The proposition Tuesdays are nath is true,” and so since I correctly believe that her and my sentences ‘Tuesdays are nath’ are synonymous, I judge her sentence true. Like her, I find myself at a loss to say whether, as I put it, Sundays are nath.

Now in all this, as far as I can see, Horwich’s theories of meaning and truth allow no ground for criticism. My friend’s term ‘nath’ has a BAP, and when I test directly her beliefs for their truth, I come out agreeing with her and agnostic just when she is. There seems to be something wrong, though, with taking seriously my new question of whether, as I put it, Sundays are nath. If you come to understand the question, you too shouldn’t take it seriously as something unknowable. Horwich’s theory needn’t tell us why, but it should at least leave room for thinking Nancy’s concept somehow lacking, and I don’t see how it does so. (Maybe the term is vague, and Horwich does devote a chapter that I won’t discuss to vagueness. I’ll simply report, without further justification, that I haven’t found the chapter to resolve my worry.)

2. The Meaning Role and Basic Acceptance Properties

One aspect of Horwich’s theory could hardly be wrong, and he shares it with many theorists with whom he strongly disagrees on other matters. In everyday thought, at least, meaning figures in explanations of things people do and say. That leaves the question of whether the concept meaning—or some approximation to it or precisification of it—would figure in the best scientific accounts of us. Suppose, though, fantastically, that we had a full scientific account of everyday human phenomena, and that something in the account plays pretty much the role that talk of meaning plays in our everyday explanations. Then whatever plays this meaning role in the theory is meaning. Quine in effect proceeds from this dictum, but then finds that nothing plays the meaning role very well, and many alternatives play it more or less. Horwich thinks that meanings, as the dictum would reveal them if we knew enough, are fairly determinate.

Our first problem, in proceeding from this dictum, will be to identify what the meaning role is in an explanation of human phenomena. The next problem will be to see whether anything in human affairs plays this role. Of course we don’t need to answer this second question for any conceivable theory of humanity, if we can identify features that a correct theory of us is bound to have—and we can hope to learn much from good, informed conjectures about aspects of a correct theory of human phenomena. Horwich proceeds to make some conjectures, and then works to identify what the meaning role would be in a theory of humanity that fits these conjectures.

Horwich’s central conjecture is that the deployment of words in speech and in thinking can be explained in a particular kind of way. For each word w, Horwich conjectures, there is a non-semantic feature that explains its overall deployment. This feature, moreover,
will be an acceptance property of the following form:—‘that such-and-such \( w \)-sentences are regularly accepted in such-and-such circumstances’ is the idealized law governing \( w \)’s use.\(^{12}\)

These basic acceptance properties, which I am calling “BAPs”, can be quite various, according to Horwich, such things as (roughly) that ‘That is red’ is accepted for red things, that ‘The bachelors are the unmarried men’ is accepted always, and instances of the schema ‘\( \langle p \rangle \) is true iff \( p \)’ are accepted. Of course no such property explains the word’s deployment all by itself; rather, the BAPs for various words combine with other aspects of psychology and context to do the explaining. (Thinking and uttering in words Horwich treats in this same way, with thought in effect being like utterances to oneself, whether in one’s spoken language or in a language of thought (9).) Horwich’s basic conjecture is that explanations that take this form will serve to explain human thought, utterances, responses to utterances, and the like, and that the intuitive role of meaning and thought content in common sense explanations will be played by BAPs.

Of course we don’t fit these regularities without exception, and this leads many theorists like Brandom and Kripke’s Wittgenstein to reject all such regularity theories. Horwich has a response: these are ideal laws, familiar from almost any of the sciences—as with ideal gas laws, for instance. There’s no special problem on this score for a theory of meaning.

Most of the things we accept don’t have the observational or axiomatic flavor of “That’s red” or “Bachelors are unmarried men.” How does Horwich envisage explanations of deployment as working. A belief that North Korea won’t give up its nuclear weapons might explain a diplomatic maneuver; how will the BAPs of \textsc{give up, nuclear, weapons} and the like figure in such an explanation? Horwich doesn’t say much on this score, and a full answer might be intractably complicated. To evaluate Horwich’s conjecture, though, we would need to get some idea. Horwich takes modus ponens as one example of how the acceptance of complexes can be explained. How might this work? Approaching a traffic light, say, I accept, “If it’s red, it will soon turn green.” I accept this on the basis of induction or something of the sort, which is a substantive epistemic matter not explained by meanings alone. On seeing the light, which is red, I accept ‘It’s red’ and apply modus ponens, accepting ‘It will soon turn green.’ There are many complexities here that a good artificial intelligence model would have to accommodate, but I’ll focus on \textit{modus ponens}. The sentence ‘If it’s red, it will soon turn green’ is built up by conditionalization from ‘It’s red’ and ‘It will soon turn green.’ The BAP that gives meaning to the word ‘if’ (or more

\(^{12}\) P. 28 and elsewhere. This formulation is repeated many times to accompany further elucidations (Each time an unintended ‘is’ comes at the end, which I omit.)
generally, to the device of conditionalization) is that one who accepts the conditional and the antecedent accepts the consequent. This law explains my accepting the consequent.

Horwich’s central conjecture is that more complex explanations along such lines can be given for any regular, lawlike aspect of the deployment of language and formation of beliefs, and that action explanations that invoke beliefs will have BAPs figuring in them in parallel ways. Of course some behavior will constitute exceptions to the ideal laws that figure in such explanations, but that’s in the nature of idealizations, Horwich insists.

Whether this conjecture about the form that explanations in human affairs will take is plausible, given our current knowledge, I’m not qualified to judge. The best imaginable confirmation would be an artificial intelligence model in a robot, a model that took the form Horwich proposes and accounted for those cases of sentence-acceptance and action that common sense expects. My understanding is that artificial intelligence models have turned out to be full of pitfalls, with everyday human common sense, however dependable it may be, proving frightfully hard to simulate. Computer scientists speak of the “frame problem”, and for Horwich’s program to work, it would have to include a solution to the cluster of problems that go under that term. In particular, we don’t know whether common sense thinking can be understood as consisting in syntactic operations of a language of thought, as Horwich’s chief example modus ponens seems to suggest. I would think that in the face of these things that we don’t understand, we must be cautious in our predictions of what kinds of explanations of the deployment of words will turn out to work. A solution might reveal what meaning consists in, but a solution is far from at hand.

The BAP of a word depends on other words, and reviewers of Horwich’s earlier book Meaning doubted that we could specify a BAP that the words ‘dog’ and ‘chien’ have in common. A Frenchman’s acceptance of ‘C’est un chien’ in certain conditions $D$ constitutes the meaning of ‘chien’, and an Englishman’s acceptance of ‘That’s a dog’ in those same conditions $D$ constitutes the meaning of ‘dog’—the very same meaning. How, though, can we understand these as the same property? The matrices ‘That’s a …’ and ‘C’est un …’ are parallel in the two languages, and we might speak of the meaning matrix that is a … that the two have in common, composed of that, is, a, and a like sentence structure. The BAP of the word ‘dog’, then, is that a linguistic construction meaning that is a …, applied to this word, is accepted in conditions $D$. This works if the French construction ‘C’est un’ also means that is a …, in that ‘Ce’ in French and ‘That’ in English share a BAP and likewise for BAPs is, a, and the structure in which they combine. It’s fine, according to Horwich, for the BAP of ‘dog’ to depend on the BAPs or other words such as ‘that’. How, though, the worry now arises, do we specify the property that that English ‘that’ and French ‘ce’ share. There’s a parallel, to be sure, and if we can specify the parallel, we can specify the meaning of English ‘that’ and French ‘ce’. The parallel, though, seems
to involve words like ‘dog’, ‘cat’, ‘is’, and ‘a’ in English and their translations in French. Horwich allows mutual dependence of meanings, but he wants dependence to be mostly local and mostly one-way. For framework words like ‘that’, ‘is’, and ‘a’, critics haven’t been able to see how he can have this. Isn’t meaning going to turn out to be holistic, as Quine insisted? For a word like ‘dog’, Horwich may not be in trouble, so long as he can have the meanings of framework words like ‘that’, ‘is’, and ‘a’, but it is hard to see how to specify the meanings of these framework words except by their interaction with the words that they will later be used to explain. There may be nothing wrong with this, but it doesn’t fit Horwich’s anti-holistic conjecture—or if it does, I don’t see how.

Despite all this, a conditional claim must surely be right: that if for each word, a unique BAP explains its deployment, and if the role that BAPs play approximates the role that common sense ascribes to meanings, then the meaning of a word is its BAP, and the meaning of a sentence is the structure of BAPs that make it up. After all, it seems truistic that if something unique plays the meaning role in human explanations (whatever that role turns out to be), that’s meaning. That leaves the questions of whether Horwich has correctly identified what the meaning role is—or whether anything unique, or close to unique, plays that role.

In the rest of this essay, I mostly put aside these broad doubts, and try to follow out the consequences of treatments that Horwich himself gives. Even if Horwich’s lines of analysis are fairly successful, I’ll argue, they leave room for concepts of meaning that Horwich thinks he can refute.

3. Ideal Laws

Horwich must be right that idealization plays a legitimate role in explanations through meaning, and that this is of a piece with idealization in a wide range of other scientific explanations. Horwich speaks, though, as if there is such a thing as the ideal law that governs the deployment of a word. That doesn’t fit the way idealization works in other sciences. Horwich cites as a parallel ideal gas laws, but a gas can be idealized in more than one way. To a first approximation, a gas consists of point masses. To a second approximation, it consists of spheres of a certain diameter. Each of these models explains much. With either one, the gas behaves roughly as it would if it fit the model precisely, and it does so because its main tendencies are modeled. For such an explanation to work, there needn’t be a unique model with this virtue. There may even be—as Quine thought for meaning—a vast number of models that can serve to explain.

13 I’ll ignore differing degrees of freedom.

Horwich takes the idealizations that figure in meaning theory to be of a piece with the idealizations of gas law. Meaning theory, though, faces special burdens in distinguishing the lawful regularities that constitute meaning from lapses and the like—Chomsky’s “competence” versus “performance”. Horwich might be helped, on this score, by a resource he thinks he doesn’t need: biological, evolutionary thinking. As Millikan stresses, meaning as we know it is a biological phenomenon: only living beings and their products literally mean anything. If the deployment of words works as Horwich conjectures, that is presumably because of a history of genetic selection for propensities to language. Recognizing this is consistent, to be sure, with Horwich’s approach as it stands, since it allows that what was selected for was exactly the kinds of propensities that Horwich describes, which are specified without bringing in talk of biological function. But Horwich specifies these as ideal laws, and the sense of ‘ideal’ that he needs might be specifically biological. Concepts like \textit{muscle} involve common sense teleology: wings are for moving by contraction and relaxation. In biology this cashes out in terms of histories of selection. The as-if purposes of biological function distinguish normal, proper functioning of a mechanism from defects. The ideal features are the ones that are there because they contribute to the mechanism’s as-if purposes. Heart disease isn’t explained by an ideal heart but by deviations from the ideal, and the best explanations both of normal and of abnormal functioning of a heart may single out the ideal heart that does what hearts have been selected genetically to do: pump blood effectively. Meaning may likewise pertain to what’s ideal in this biological sense, and not the language faculty’s defective operations.

However we characterize what’s “ideal” in the sense the Horwich needs, a legitimate thought can then drive an approach like Horwich’s. Any satisfactory idealization of the deployment of words, we might conjecture, will attribute meanings to them—meanings in the sense of properties that play the meaning role. Or perhaps some class of such models attributes meanings. Meaning, then, is whatever plays the meaning role in those good ideal models of the deployment of words in which something plays that role. Horwich asserts this, and adds the conjecture that the meaning role takes a certain form, the combining of BAPs.

Once we put matters in such a way, we can see what are the conceptual and empirical burdens that a straightforward theory of meaning shoulders. It requires specifying what the meaning role is. It requires too that for each word, some unique property plays that role. A less straightforward theory of meaning might specify what synonymy consists in

---

15 The concepts of possessing a word and accepting or rejecting a sentence might perhaps best be explained in terms of what these states are “for”, although Horwich himself sketches an account that makes no appeal to teleological considerations (\textit{Meaning}, 94–95). My point here is that even if these notions were specified in teleological terms, what Horwich does with them might not require further teleological thinking.
without saying anything informative about what property synonymous expressions share. This would require identifying the synonymy role, and would demand that some unique relation between expressions play this role.

Quine worked with synonymy, and thought that once we say all we can about the synonymy role, we’ll find that many alternative relations play that role. Horwich in effect conjectures that at least one good model of the deployment of words has a meaning role—so that synonymy is sameness of what plays this role. He conjectures that for a given word, the same property plays that role in any good model that has a meaning role for that word. Or at there is a close approximation to such determinacy.

4. Compositionality Narrow and Broad

Many theorists take compositionality to be a feature of thought and language that cries out for explanation and sharply constrains the form that any explanation of thought and language can take. Davidson takes explaining compositionality to be a chief virtue of his truth- and reference-theoretic treatment of meaning. Horwich, though, devoting a chapter of *Meaning* to the subject and entitling the final chapter of *Reflections* “Deflating Compositionality”, asserts, “Davidson’s problem (of how we might derive interpretations of complex expressions) has a trivial solution” (202f). Compositionality comes for free, in that the meaning of a complex structure just is its compositional structure—the way it is composed from elementary words with certain meanings.

This piece of deflation reads “compositionality” narrowly. “Compositionality”, as Horwich means the term, is strictly this: that the meaning of a linguistic composite is fixed by the meanings of its parts and how they go together. This specific demand, Horwich is right, can be met trivially and by fiat: let the composite’s meaning just be a structure composed of the meanings of its elements.

Still, we might read the demands of “compositionality” in a broader way. Horwich first puts these demands as two questions a theory must answer:

What form would our hypotheses about the meanings of someone’s words and sentences have to take in order that the latter be deducible from the former? And for that matter, how does the understanding we have of our own language derive from our understanding of its basic elements? (198).

The first of these speaks to compositionality in the narrow sense, but the second is broader. Beth asks me, “Could you bring me the yellow towel that is stacked at the back of the closet at the end of the hall?” I bring it to her and she smiles—and these happy results ensue because I have understood her. Phenomena like this strike me as prime indications of compositional structure, indications that cry out for explanation. How did she and I accomplish her wishes? It’s an empirical fact that things like this happen; they are
prime examples of what Horwich calls the “deployment” of words. The vignette manifests compositional understanding, and in a broader sense, the problem of compositionality is to explain phenomena like this.

Horwich in no way purports to deflate compositionality in this broader sense. What he does is to presuppose it. He presupposes, in his account of meaning, that phenomena like these will have explanations, and that the explanations will take a particular form. The BAPs of words do all the explaining of regularities in the deployment of words, he conjectures—or more exactly, all the explaining that stems from meaning. BAPs explain by applying recursively. If no properties are capable of doing this, then, on Horwich’s account, there are no such things as meanings. The very existence of meaning, then, depends on BAPs’ working recursively in explanations of sentence deployment. Compositionality in the broad sense is thus built into Horwich’s whole project. As for the narrow sense, it may be trivial that if meanings exist, they are compositional, but that compositional meanings exist is anything but trivial. That they exist amounts to saying that the composition of BAPs explains, along with non-meaning facts, whatever is lawlike in the deployment of language.

Horwich says this, to be sure, but he insists that compositionality is another matter. “The constraint we need in order to obtain a good theory comes, not from the compositionality of meaning, but rather from the use import of meaning, namely, that the overall use of a complex is explained, in part, by the meanings of words and how they are combined” (218).

As for understanding, this second part of compositionality as Horwich characterizes it deflates away too. Grasping a sentence consists in no more than understanding its elements and how they combine. The understanding in question, though, is implicit (219–20), and a lot is packed into such talk of “implicit understanding”. What’s packed in, indeed, is the very sorts of things that enter into explaining the vignette—things that no one would claim could be deflated away.

Horwich’s talk of “deflation” is of course aimed at particular sorts of arguments in the literature, arguments that exploit a demand of compositionality to rule out all theories but one. Whether we should count recursive explanations of sentence use as displaying “compositionality” won’t bear on whether Horwich succeeds in demolishing these arguments. Either way, Horwich offers an alternative to Davidson: to explain recursively not truth but sentence use. Whether either approach works is daunting to assess, but each acknowledges compositionality both narrow and broad.

Anil Gupta, in his review of Horwich’s Meaning, argued against this aspect of Horwich.16 Whereas truth is well-behaved, he argued, use is intractably messy. Horwich

---

16 Gupta reference xxx.
proposes to cope with the mess by idealizing, but whether or not this works, the decorum of truth conditions gives no overall advantage, it seems to me, to meaning as truth-conditions. Truth and ideal use are explanatorily on a par. Semanticists’ evidence, after all, consists largely of answers to questions like this: “Imagine a floating log moves with the current to a position under the bridge. Can we say, ‘The log floated under the bridge’?” Evidence for truth-conditions thus consists in evidence that two sentences are accepted alike: a test sentence (like “The log floated under the bridge” under one reading) and a long-winded but unproblematic description of a situation. Elicited responses to questions like this yield truth conditions—but not invariably: that they do amounts to an ideal law. The same need for idealization, then, intervenes between empirical data and meaning on either approach, and truth conditions amount to an aspect of ideal use. Truth conditions and ideal use alike fit neatly with meaning—and with experience messily.

Horwich, as I say, hypothesizes a substantive, empirical explanation of broad compositionality, an explanation sharply different from Davidson’s. What composes, Horwich conjectures, isn’t reference and truth conditions, but BAPs. Now it does seem a safe bet that the explanation of phenomena like my finding the towel that Beth wants will be somehow recursive. That it will take the form that Horwich conjectures strikes me, in contrast, as a bold and risky bet. (Or it would be bold if there were great penalties in philosophy for being wrong.) Ultimately, the test would have to lie in computer programs of artifical intelligence that filled out Horwich’s conjecture and succeeded in capturing everyday phenomena that we would regard as displaying normal good sense. The phenomena of language and semantics (on a broad understanding of the term) have proved frightfully difficult to handle theoretically. Perhaps this is because so much of the work has focused on truth conditions, taking a form quite different from the one that Horwich recommends. I’d expect, though—in my ignorance of the field—that many of the phenomena that give Davidsonian semanticists trouble would challenge Horwich’s program too, and it is far too early to say what the upshot would be. Horwich’s approach has the advantage that it doesn’t appeal to mysterious and unexplained notions of truth and reference, but we need to see more. A good first step might be very simple examples worked out for toy situations—adverbs like ‘quickly’, for instance, or the reference of pronouns. We could then ask ourselves what the meaning role is in these explanations. Can it be played by compositional structures of “basic acceptance properties” in Horwich’s special sense? Will a unique assignment of such properties to words do the job (or an assignment that is close enough to unique)? I don’t know, but these questions are highly substantive.

17 I recall the syntactically ambiguous sentence ‘The log floated under the bridge’ from a talk by James Higginbotham; Italian, he said, doesn’t permit a corresponding ambiguity.
In a rough sense they are empirical, but they will require a lot of ingenious analysis. If Horwich is vindicated, what’s vindicated will be far from purely deflationary.

5. Holistic Logical Charity

What makes a property the BAP of a word, its unique basic acceptance property. The property is that one is prone, under certain conditions, to accept certain sentences that contain the word—and perhaps to reject others. This property joins the BAPs of other words, as well as factors other than meaning, to explain all that is lawful in the deployment of the word. Crucially for what I’ll be saying later, the acceptance must be underived.

As it turns out, though, Horwich sets a further important requirement on BAPs. To see this, consider first Prior’s miscreant connective ‘tonk’. The purported laws that characterize it are these: (i) from $p$ one infers $(p \text{ TONK } q)$, and (ii) from $(p \text{ TONK } q)$ one infers $q$. This is perverse, because one is then prone to infer any conclusion $q$ from any premise $p$. TONK, some authors conclude, is a defective concept. Not at all, says Horwich; there’s just no such concept as TONK. The purported regularities, far from being the BAP that characterizes a defective concept, are impossible. They entail, for instance, accepting both a claim and its negation, and that violates the BAP of the concept NOT. Since no such regularity is possible, there could be no such BAP.

On this Horwich must be right. He goes further, though: whereas other theorists, some of them, maintain that concepts may be defective empirically or even logically, Horwich verges on claiming that there can be no such thing as a defective concept. Logically milder examples of defective concepts, though, seem at least to be in the running. Take Horwich’s example ‘Pom’, used by many Australians for Englishmen stereotyped as pretentious. On a treatment that Horwich rejects, the BAP for the concept POM is this: that (i) from $X$ IS ENGLISH one infers $X$ IS A POM, and (ii) from $X$ IS A POM one infers $X$ IS PRETENTIOUS. This, Horwich says, is not a possible ideal law of usage. What, after all, is a user of ‘Pom’—call her Tayla—prone to accept if she encounters Sasha who is manifestly English and manifestly not pretentious? Her sentence ‘Sasha is English’ doesn’t mean SASHA IS ENGLISH unless she is prone to accept it in those circumstances, and her sentence ‘Sasha is pretentious’ doesn’t mean SASHA IS PRETENTIOUS unless she is prone to reject it in those circumstances. So the lawlike dispositions that are the purported BAPs of ‘Pom’, ‘English’, and ‘pretentious’ just aren’t compossible, Horwich maintains. Conditions (i) and (ii) don’t give a possible law that could govern a person’s usage of a term.

\[18\] I am told that in fact, ‘Pom’ users don’t count an Englishmen as a “Pom” unless he has a posh manner.
Do they not? The laws in question, remember, are ideal laws that admit of exceptions. Tayla the user of ‘Pom’, imagine, in her limited experience and caught up in the attitudes of her community, might find a non-pretentious Englishman virtually unthinkable. She’d find such a phenomenon hard to wrap her mind around, and fall into linguistic confusion. Why not say that conditions (i) and (ii) indeed give the BAP of the concept POM, but that when they collide with the BAPs of ENGLISH and PRETENTIOUS, the situation is anomalous, and she deviates from the ideal laws that apply?

A stronger example still, which Horwich takes up, might be with the word ‘true’. Tarski thought that the liar paradox shows that the ordinary notion of truth is inconsistent and in need of replacement. Horwich’s theory of truth is unusual in that it doesn’t focus on semantic paradoxes or take resolving the paradoxes as touchstones of success. Now of course the regularities that lead to paradox cannot conceivably be followed uniformly. They lead to accepting both a sentence and its negation, contrary to the BAP that characterizes negation. But a law, Horwich tells us, can explain the deployment of a word even if conformity isn’t invariable. The departures Horwich has in mind are explained by such things as fatigue, inattention, distraction, and capacity limitations, and dismissed as lapses. Why can’t another source of departure be baffling surprises?

Horwich himself demands, in effect, that the BAPs that constitute the meanings of words be ones that those words could conceivably have had jointly and invariably. This fits, in a way, the general vision of assimilating explanations in terms of meanings to explanations by ideal models in science more generally. Ordinarily, the characterization of an ideal model is meant to be consistent: the specifications of the model could characterize a system, though we don’t expect that the system will fit the specifications precisely. (Whether, say, quantum theory gives rise to exceptions I can’t tell you.) Horwich might have a choice between two versions of his ideal use theory of meaning. One is global: the meaning of a word is the BAP that forms part of an assignment of a BAP to each word that the words of the person could jointly have had exactly, and whose approximate realization explains her actual deployment. This is the version that Horwich adopts. Alternatively, though, he could leave out the requirement that BAPs must be composable.

If we focus on explaining the deployment of individual words like ‘Pom’ or ‘true’ one by one, the latter might be the better choice. The naïve reaction to a semantic paradox is bafflement, and even sophisticated treatments don’t entirely remedy the problem. It is important to search for consistent rules for deploying a concept like TRUE, as Tarski, Kripke, Gupta, and others have done, but perhaps such treatments don’t describe the concept we already have but tell us how to reform it. Horwich, though, in effect demands a kind of charity in his ascription of BAPs: taken together, they must be properties that one’s use of words could have had precisely. “The fact that there are instances of the naïve
truth schema which we will (or would) not be prepared to accept,” he writes, “shows that
the simple rule (to accept all instances of the schema) is not the one that we implicitly
follow” (154n19). But when the naïve thinker encounters a paradox and is baffled, how
do we explain the bafflement? It doesn’t seem that she has lapsed from some pattern
that would never have brought her to bafflement if she hadn’t deviated. Which of the
sophisticated theories of truth now on the table shall we say she lawfully almost fit? Why
treat her bafflement as anomalous, when it was fully to be expected because of regularities
that seem closely related to the meaning of ‘true’? It seems rather that certain near-
regularities have collided, and call for rethinking—and regularities in the deployment of
true brought about the collision. Horwich, though, is not looking just for laws that we
follow pretty much, but for an entire system of laws that we could have followed completely.

6. Theoretical Terms

For many purposes, a theory that attributes a consistent ideal pattern has advantages,
and we should probably follow Horwich in demanding that the ideal laws that constitute
the meanings of a person’s words must be consistent with each other. Horwich, though,
has a reason beyond this for rejecting BAPs that could make for defective concepts. The
implicit acceptance that figures in a BAP must—as I mentioned before without comment
or stress—be entirely underived. Tayla the ‘Pom’ user, true enough, accepts the sentences
‘Englishmen are Poms’ and ‘Poms are pretentious.’ But these acceptances are rooted in the
view that Englishmen are pretentious, and if asked, she would take herself to have grounds
for this belief. She doesn’t, then, accept Pom theory underived. (Horwich’s treatment of
this in Meaning I found unconvincing, but on reading Reflections, I am convinced.)

The term ‘Pom’, we could say, is theory-laden. Horwich proposes a way to handle
theory-laden terms, a way that factors out what is underived and hence could be meaning-
determining. Tayla accepts what I grandiosely call Pom theory:

\[
\text{Englishmen are Poms & Poms are pretentious.}
\]

The meaning of her word ‘Pom’ is rooted in this theory so couched, and I’ll call this
the root theory for the word ‘Pom’. Tayla accepts this root theory on the basis of, say,
encounters with Englishmen she found pretentious, or television depictions, or just hearsay.
She doesn’t accept it, then, underived. Pom theory as a whole thus cannot be what purely
determines her meaning for the word ‘Pom’.

Rather, says Horwich, we can factor Pom theory into two parts. One part can be
put without using the word ‘Pom’ and so is independent of what Tayla’s word ‘Pom’
means. The other part is accepted underived, and so meets a requirement for being purely meaning-determining. The first, substantive part of Pom theory is its Ramsey sentence,

$$\exists P (\text{Englishmen are } P \text{'s } \& P \text{'s are pretentious.})$$

(1)

This amounts to the theory that Englishmen are pretentious. What is meaning-determining is a conditional that hedges Pom theory with this Ramsey sentence—in effect:

$$\text{Englishmen are pretentious } \rightarrow (\text{Englishmen are Poms } \& \text{ Poms are pretentious.})$$

(2)

In *Meaning*, Horwich calls this the *Carnap conditional* for Pom theory.¹⁹ The Carnap conditional will be purely meaning-determining if (i) Tayla implicitly accepts it underived, and (ii) this explains all that’s lawlike in her deployment of the word.

Presumably we could give the same treatment to the concept *true*. The root theory consists of all instances of the deflationary schema, such as

**SNOW IS WHITE** is true just in case snow is white.

The user of ‘true’, though, doesn’t accept this underived, even implicitly. He implicitly assumes that there is a property that fits this deflationary root theory for ‘true’. The purely meaning-determining sentences for ‘true’ are instances of the deflationary schema, but hedged with a Ramsey sentence saying that there is a property that universally fits the deflationary schema. Once we encounter Tarski’s paradox, we reject this Ramsey sentence. The fact that we mean *true* by ‘true’ then has no entailments for whether we will accept or reject sentences like ‘It’s true that snow is white’ or ‘It’s true that snow is black.’

Back, though, to ‘Pom’. In what sense does Tayla implicitly accept the Carnap conditional (2)—and how does this explain her deployment of the word? She hasn’t thought about Ramsey sentences and hedging; it’s *we* who do that. She would accept Pom theory as couched with the word ‘Pom’, and think herself to have a basis for what she believes. Her acceptance of Pom theory, then, can’t be what plays the pure meaning role in her deployment of her word ‘Pom’, because it can’t count as “accepted underived”. And it is much *as if* she explicitly accepted Pom theory’s Ramsey sentence (1) and its Carnap conditional (2). Whatever indications she thinks to support the tenets of Pom theory she’d think to support its Ramsey sentence if she understood it, and she would regard the Carnap conditional as empty, not as the sort of thing that could be supported by evidence. We attribute a BAP to her word ‘Pom’ as part of a rational reconstruction

of her use of the term. It plays a pure meaning role in an ideal model that goes to explain her deployment of the term.

The Carnap conditional for a theory-rooted word thus qualifies as a kind of sentence whose implicit acceptance underived could play the meaning role in explaining its deployment. The explanation takes the form of an ideal model: the user reasons for the most part as if she accepted the conditional underived, conformed to ideal laws governing other terms, and formed beliefs in certain ways. [Verbose xxx?]

7. Indeterminacy

All this still leaves us with the question of determinacy: how determinate will meanings be on this approach? Suppose, as Horwich conjectures, a person’s deployment of a word can be explained by her underived acceptance of certain sentences that contain the word. Still, alternative good explanations might invoke different sentences in this role, yielding different attributions of meaning. A little indeterminacy of this sort would fit into a common-sensical conception of meaning, whereas vast Quinean indeterminacy would not. Quine thought meaning, on the best sense we could make of it, to be vastly indeterminate, and Kripke, speaking for Wittgenstein, thinks it vastly indeterminate so long as we think there is some natural fact of the matter as to what we mean by our words. (He thinks too that a more determinate conception of meaning must be normative.) Horwich thinks meanings to be ordinary matters of scientific fact. The possibilities are narrowed down, he argues, by the requirement that meanings must explain in the right way. As I have been putting the requirement, they must play the meaning role in explanations of human affairs. How much does this reduce indeterminacy? To inquire into this, I take up a stock example from science.

Consider the terms ‘mass’ and ‘momentum’ (written ‘m’ and ‘p’) in Newtonian theory. Suppose, for the sake of simplicity, that Einsteinian special relativity has it right, and that you and I speak its language. Call our Newtonian ‘Newtonia’. Special relativity holds out a number of possibilities for what these terms could mean in Newtonia’s mouth. For ‘mass’ there are rest mass $m_0$ and relativistic mass $m$. In Einstein’s first formulation there were also longitudinal mass $m_l$ and transverse mass $m_t$. Perhaps the Newtonian’s ‘mass’ means one of these—or perhaps Kuhn is right that Newtonian concepts are incommensurable with Einsteinian ones, and the Newtonian’s term ‘mass’ means none of these things. How will Horwich’s theory pronounce on all this? Will it yield a unique ascription of Newtonia’s meanings?

Newtonia accepts that “mass” is constant. (She can put this as “$\dot{m} = 0$,” where a dot like this one indicate rate of change with time.) We ourselves think this to be correct so long as “mass” is rest mass $m_0$, but not if it is relativistic mass $m$, longitudinal mass $m_l$,
or transverse mass $m_t$. She accepts too a law relating force $F$, “mass”, and acceleration $\dot{v}$:

$$F = m \dot{v}.$$  \hspace{1cm} (3)

(I’ll take ‘$F$’ to mean force in both her mouth and ours, and ‘$v$’ to mean velocity, so that “$\dot{v}$” is acceleration.) We think this right if “$m$” is longitudinal mass $m_l$ or transverse mass $m_t$, but not if it is rest mass $m_0$ or relativistic mass $m$. Finally, she accepts the following pair as characterizing “momentum” or “$p$” and giving a correct force law, that force is the rate of change of “momentum”:

$$p = m v, \text{ and } F = \dot{p}.$$ \hspace{1cm} (4)

We think this true if “$m$” is relativistic mass $m$, but not if it is rest mass $m_0$, longitudinal mass $m_l$, or transverse mass $m_t$. What Newtonia gets right and what she gets wrong depends on which of these four candidates, if any, is what she means by her term ‘$m$’ (and consequently, what she means by her term ‘$p$’).

Horwich’s theory of meaning should give us the answer; how does it pronounce? One way to apply the theory might be to say this: The three claims of (3) and (4), along with the constancy of “mass”, constitute a fragmentary Newtonian theory. This, Horwich’s might say, is the root theory that gives meaning to her terms ‘$m$’ and ‘$p$’. Thus to get the BAPs for these terms, we can form the Ramsey sentence and the Carnap conditional for this root Newtonian theory. Newtonia’s implicit, underived acceptance of this Carnap conditional, Horwich could then say, is the BAP constituting the meanings of her terms ‘$m$’ and ‘$p$’. Since the Ramsey sentence is false, this Carnap conditional has no logical implications for the actual, relativistic case.

Alternative BAPs for her terms ‘$m$’ and ‘$p$’ might be found, however. We could say that ‘$F = m \dot{v}$’ by itself is the root theory for Newtonia’s term ‘$m$’. To get the meaning-determining Carnap conditional, then, we hedge it by its Ramsey sentence, which is

$$\exists x (F = x \dot{v}).$$ \hspace{1cm} (5)

This Ramsey sentence drops the term ‘$m$’ and says that something is systematically related to force and acceleration as $m$ is in the formula “$F = m \dot{v}$’. Independently of the meaning of ‘$m$’ in her mouth, Newtonia implicitly accepts this. The purely meaning-determining sentence for her term ‘$m$’ is thus, on this analysis, the Carnap conditional, namely the root theory for ‘$m$’ hedged by this Ramsey sentence (6). This gives us

$$\exists x (F = x \dot{v}) \rightarrow F = m \dot{v}.$$
Her term ‘momentum’ or ‘p’ can then be given a similar treatment, with its root theory being ‘\( p = mv \)’, and its meaning-determining sentence thus being the Carnap conditional,

\[ \exists x (x = mv) \rightarrow p = mv \]

This analysis reads Newtonia’s “\( m \)” as, if anything, Einstein’s longitudinal mass \( m_l \) when the acceleration is in the direction of the motion, and as his transverse mass \( m_t \) when the acceleration is transverse to the motion. What she calls ‘momentum’ or ‘\( p \)’ isn’t momentum, and whereas she is right that, as she puts it, “\( F = m\dot{v} \),” she is wrong that “\( F = \dot{p} \),” and wrong that “\( \dot{m} = 0 \).”

A third alternative has Newtonia speaking our own language, couching a mixture of true and false beliefs in that language. This time let the root theory for ‘\( m \)’ and ‘\( p \)’ be \( (4) \), which, recall, is ‘\( p = mv & F = \dot{p} \)’. The Ramsey sentence for this is

\[ \exists x \exists y (y = xv & F = \dot{y}) \]

Hedged by this Ramsey sentence, the root theory \( (4) \) becomes the Carnap conditional

\[ \exists x \exists y (y = xv \& F = \dot{y}) \rightarrow (p = mv \& F = \dot{p}) \]

We now say that she speaks our language, rightly believing that \( p = mv \) and \( F = \dot{p} \), and wrongly believing that mass is constant and that \( F = m\dot{v} \).

We thus have three different ascriptions of meaning to Newtonian’s term ‘\( m \)’. Each ascription has her accepting the four claims of Newtonian theory that she accepts—namely \( (3) \), the two claims of \( (4) \), and the constancy of “mass”. Is there a fact of the matter which one of these interpretations is right? We might try to settle the issue by asking in what terms she would couch relativity theory. But if relativity theory were put to her she would find it confusing, and once she came to accept some formulation of it, we would have to ask whether she had changed her meanings. That amounts, on Horwich’s official theory, to asking what Carnap conditional for relativity theory she implicitly accepts already—implicitly accepts underived. But what she implicitly accepts underived isn’t a straightforward matter; it is a matter of what ideal models explain her deployment of terms. If she accepts all of Newtonian theory as a matter of habit, each of our three candidate idealizations may equally explain her deployment of the terms in the theory. We are still left with the question, then, of how to go from her dispositions to deploy her terms ‘\( m \)’ and ‘\( p \)—her actual dispositions—to an ideal explanatory model of these dispositions. In this case, alternative models all explain her deployment of terms, and they differ in which Carnap conditionals they have her accepting.

Gibbard, March 6, 2007
Like things hold if we ask what she would say if she came to doubt the constancy of “mass”. Eventually, perhaps, she would come to a formulation of her views but we would be left with the question of whether the meanings of her terms had then changed. Again, there might well be no fact of the matter which Carnap conditional she implicitly accepts underived, and so no fact of the matter what her terms ‘m’ and ‘p’ mean.

What I have done, in effect, is to test whether Horwich’s strategies yield determinate meanings by applying his account, in a Quine-like way, to a fragment of Newtonian theory. Which parts of the fragment are meaning-determining has turned out to be indeterminate—and what’s true and what’s false in the fragment partly hinges on the meanings that are indeterminate. Determinately, only the theory as a whole is true or false, but indeterminately, parts of it might be.

8. Analyticity

With determinate meanings come facts of the matter as to which claims are analytic and which parts are synthetic claims of substance. If meanings are utterly indeterminate, the analytic-synthetic distinction collapses, so that nothing is then determinately analytic or determinately synthetic. Skepticism over analyticity stems from skepticism over determinate meanings. Or so we might think. Horwich’s himself, though, while claiming a high degree of determinacy for meanings, rejects claims traditionally made for analytic or conceptual truths. He debunks claims that analyticity can ground our knowledge of a priori truths. His reasoning is subtle, and I’m not clear precisely what his conclusions are. Still, I'll try to explore how analyticity would stand if meanings were as determinate as Horwich thinks they are.

The word ‘bachelor’, we might have thought, means UNMARRIED MAN (in the sense of a man who has never been married). Not so, according to Horwich: ‘bachelor’ is a single word, and so has as its meaning its BAP, whereas ‘unmarried man’ is a phrase whose meaning is a complex of at least two BAPs. The tie of the word ‘bachelor’ to being an unmarried man, according to Horwich, is rather this: the BAP of ‘bachelor’ is that one accepts, underived, the sentence, ‘The bachelors are the unmarried men.’

Is this sentence, then, analytic? Or equivalently, is the proposition THE BACHELORS ARE THE UNMARRIED MEN a conceptual truth? Horwich says no. Anyone who understands the sentence, to be sure, is prone at least implicitly to regard it as true. (That is to say, the ideal model in which her meanings figure has her accepting the sentence.) Anyone able to entertain the proposition—including people who don’t speak English but can express the proposition in their own language—is similarly prone to regard it as true. Still, we can’t deduce from any of this that the proposition in fact is true.
If the proposition isn’t true, then anyone who entertains it must, implicitly at least, be making a mistake. How can this be? The most direct way to test whether a proposition is true or false is to entertain it and determine whether that’s the way things are. We can test the prediction that it will rain tomorrow by waiting until tomorrow and seeing if it rains. Now anyone who conducts such a test of a meaning-determining sentence will conclude that it passes the test—on pain of deviating from the ideal model of her use of words. If she rejects the meaning-determining sentence, then either she fails to possess the concept, or she is somehow deviant: her judgments are either false or non-ideal.

Could a concept be like this? Can it be that in this sense, anyone who even possesses the concept must ideally be making a mistake? Perhaps so if some concepts are “defective”, as \textsc{tonk} is said to be, so that one can’t so much as entertain propositions with the concept without making a mistake. Horwich, though, as I said, thinks that concepts can’t be defective in this way—and I argued the best version of his theory does indeed yield this consequence.

Still, Horwich is right that there is no sheer inconsistency in admitting that a sentence is meaning-determining for a term, but claiming that it is false. Rush, imagine, unable to contemplate the propositions of string theory, is in his ignorance so hostile to the enterprise as to claim that none of the central tenets of the theory are true. You show him a sentence that looks like gobbledygook to him, and tell him that it is meaning-determining for the term “string”. He believes you on this, but none the less, says he, the sentence is false. You tell him that if he understood the sentence, he’d think it true. He admits this, but in that case—he stubbornly maintains—he would have corrupted himself, and be wrong.

There’s no inconsistency in this stubborn position—or not one that I can discern, if Horwich’s account of meaning is right. Legitimate uses of words, as I said earlier, extend beyond what meaning-determining sentences entail. This applies to the concept \textsc{true}: I may reasonably think a theory true even if I am utterly incapable of understanding it. Perhaps I have good reason to think that these theorists know what they’re talking about. Or I can have good reason to think them charlatans spouting falsehoods, though again I don’t understand what they are saying. All this Horwich’s account of truth and meaning allows.

Still, a meaning-determining sentence has the special property I stated: one can’t understand it without ideally thinking it true. Moreover, as far as I can see, there will never be a good reason to think the possessors of a concept \textit{ipso facto} mistaken. There would be such a reason if “defective” concepts were possible, but they aren’t—or so Horwich has convinced me. If a sentence is purely meaning-determining and hence non-defective, to think it false would, it seems to me, be silly. It would evince a gratuitous distrust of all power of inquiry. After all, anyone who understood the sentence would ideally think it
true, and we have found no reason to think he would then be mistaken. We have found no imaginable case where it seems that to understand a sentence correctly, one must ideally make some kind of mistake.

This amounts to a transcendental argument for the truth of any purely meaning-determining sentence. To understand the sentence is to think it true, at least ideally, and so directly testing its truth will ideally convince anyone of its truth. That’s not quite how “analyticity” has traditionally been conceived, to be sure, but it is a property that should convince us that the sentence is true. The term “analytic” seems a good one for sentences with this property. And in this sense, that bachelors are unmarried is analytic, rooted purely in meanings.

If a proposition is purely meaning-fixing, ought we to accept it? One question is whether we ought even to entertain it, but given that we ought to entertain it, we ought to accept it. Here is an argument that we ought to do so: If we do as we ought, after all, there is no alternative; a purely meaning-fixing proposition isn’t one that a person could entertain and reject. To reject the sentence that expresses it—the sounds as thought to oneself—would be to give it a different meaning. The argument for this limited normative conclusion is a straight argument. It isn’t an argument that can be one’s basis for accepting a proposition, since purely meaning-fixing propositions are, if entertained, accepted underived. But it is an argument that can be followed even by someone who doesn’t herself possess the concept, even someone who thinks that one rationally ought not to possess it.20

[the above seems wrong xxx. Can I show that I can’t think I ought to reject it.]

There’s an escape to consider from this argument. The acceptance that fixes meanings and concepts is ideal and implicit. This is acceptance, I argued, in the ideal model that serves best to explain one’s deployment of one’s words. Perhaps one would be thinking truly if in one’s explicit acceptances and rejections, one departed from this ideal explanatory model. To assess this possibility directly, I make myself into someone whose deployment of words is best explained by the model, and then accept or reject the proposition in question. If I reject it, I depart from the model that best explains my overall deployment of words—but such departures happen frequently. Do we have any reason to think that in this departure, I’m doing what I ought not to do? It may seem bizarre to

20 The oughts that apply to entertaining a proposition or not are quite different from the oughts that apply to accepting it. Oughts of acceptance are epistemic; they govern belief, not choice. Oughts of entertaining are matters of how to direct one’s attention, to the degree that one has control of the matter. They are thus oughts governing the will. See my Wise Choices (1990) for a discussion of how normative judgments govern states like belief that can’t be had at will.
say that I do, but I don’t have a knock-down argument against it.  

Suppose I’m right on this, however. Even so, to his arguments against truth purely in virtue of meaning, Horwich adds that purely meaning-determining sentences must be so thoroughly hedged as to lose their interest. Take the inclusive or of classical logic and the disjunctive syllogism: \( p \lor q \) and \( \neg p \), therefore \( q \). Is this accepted purely in virtue of its meaning? No, an intuitionistic logician understands the classical or but rejects it. Any sentences that are purely meaning-determining for or, then, must be hedged in such a way that even a mathematical intuitionist implicitly accepts them. The sentences he accepts won’t, by his lights, entail the validity of disjunctive syllogism. What the classical and the intuitionistic logician accept in common, then, must be hedged by some proviso: perhaps, that there is a connective for which, among other things, disjunctive syllogism is valid. With this part of Horwich’s argument I fully agree. We can’t, then, vindicate the aspects of our reasoning that we chiefly care about—even some of the clearest parts of logic—as rooted purely in meanings.

Horwich has another objection to “semantogenetic” vindications, though, which I think requires caution. A sentence can be purely meaning-determining only if it is accepted underived. No one, then, accepts a sentence on the grounds that it is meaning-determining— for that would be to accept it on some basis. Suppose, for instance, the purely meaning-determining sentence for \textsc{bachelor} is ‘The bachelors are the unmarried men.’ (Horwich proposes this tentatively, but if it needs hedging, hedge it as you wish.) Speakers of English accept this sentence on no further basis—and in particular, not on the basis that it is true in virtue of what they mean by the word ‘bachelor’. All true enough. Talk of analyticity in the philosophical tradition, however, hasn’t chiefly been used to get people to accept sentences because they are analytic. Mostly, as Horwich says, they accept the sentences already. The point has been to mark a particular way in which accepting a sentence can be beyond reproach. Whether the bachelors are the unmarried men would be a silly target for Cartesian doubt, because that they are is meaning-determining, and so empty. This “claim” that bachelors are unmarried is hardly a claim, and is different in this regard from, say, the bases of induction.

In any case, not all \textit{a priori} truths are purely meaning-vindicable in this way: not all \textit{a priori} truths must be believed, ideally, to be understood. I agree with Horwich that

---

21 Horwich raises the possibility 164g of a possession condition for a concept that is incorrect, or of a “primitively compelling belief-forming rule” (Peacocke’s term) “that cannot be made correct by any judicious selection of referent.” This doesn’t explain “how the concept-constituting character of a belief-forming rule can be responsible for our obligation to obey it” 165c.

22 Horwich discusses the status of \textit{modus ponens}, but I myself have an easier time thinking about the example of disjunctive syllogism.
there are claims that we ought to accept on no further basis, and that many of these a person could understand and reject—and indeed fail to accept even implicitly, indeed, even implicitly. Whatever basic normative truths govern expectations for the future, for instance, are cases in point: that we ought to expect the future to resemble the past, say, or that we ought to conditionalize on such-and-such a prior array of credences. One could reject such a claim, form one’s expectations of a different basis, and still understand everything involved. As for whatever truths ground controversial aspects of logic, Horwich argues in effect that a person could understand them but reject their presuppositions—as with intuitionists who understand the classical ‘or’ but think it has no sensible application to mathematics. Still, whereas some of the things we ought to accept on no further basis are substantive or have substantive presuppositions, others are merely conceptual: they are truths that one couldn’t even entertain without—at least implicitly—accepting them. Sentences that are suitably hedged so as to be purely meaning-determining would be prime examples.

If we accept a proposition underived but then ask why it’s all right to do so, it’s one thing to regard it as a basic postulate of reason, and another to think it purely meaning-determining and hence non-substantive. Why use the past to guide expectations for the future? Justification comes to an end somewhere, we can respond, and this is where it does. That’s not a justification, but a claim that no further justification is called for. Why think bachelors to be unmarried? That’s a sheer conceptual truth, and no further question arises. Purely meaning-determining sentences are true, though my grounds for saying this aren’t a matter of straightforward derivation, but transcendental in the way I sketched earlier. The transcendental vindication won’t be anyone’s ground for thinking bachelors unmarried, but it does explain why any call for a justification would miss the point. And this is a different kind of explanation from the one that would apply to the basic oughts of induction.

9. Meaning and Normativity

We ought to think in certain ways: we ought, for instance, not to accept both a claim and its negation. This in no way shows, Horwich insists, that the concept of negation is itself a normative concept, somehow laden with ought. We can explain negation without bringing in the oughts of the matter. In rain one ought to use an umbrella, but that doesn’t show that the concept UMBRELLA is a normative concept, or that RAIN is. Likewise with the oughts of negation. Others might think that negation is characterized by its logic, and that logical relations are proprieties or oughts. Horwich, though, thinks he can explain the concept NOT in a fully scientific, non-normative way.

Gibbard, March 6, 2007
Now as I have said, if Horwich’s basic empirical conjecture about how word and sentence deployment is best explained is right, then it’s hard to see how he could be wrong about what meaning is. I have been arguing, though, that—at least within limits—Quine was right. When claims are heavily rooted in a theory that is accepted without question, Horwich’s strategies don’t separate out which sentences are meaning-determining and which substantive. Alternative ideal models may equally well explain the deployment of words in the theory, and they will differ in which claims of the theory come out as meaning-fixing.

[Indeterminate baps give indeterminate meaning? Or, where there’s indeterminacy, there’s room for more determinate concept of meaning xxx.]

This leaves an empirical slack that oughts might fill. The oughts in question won’t be moral or prudential: if we ought morally to accept the entailments of the things we accept, that’s a substantive moral claim, not a sheer matter of meanings, and if we ought prudentially to accept them, argument and evidence are needed. Still, there’s an epistemic sense of ‘ought’ in which we ought to believe what our evidence supports, and in which we ought to believe the entailments of the things we ought to believe. Working out just what this sense is would be a project, but the meaning of a term, we might think, has much to do with what it entails and what entails it. With an empirical term, it has to do with what counts most basically as evidence—and the evidence for a claim, we might think, means something like that which we ought to weigh toward accepting the claim. Meanings explain usage, but if the causal-explanatory demands on meanings leave slack in determining what terms mean, perhaps the oughts of the matter fill some of the slack.

When Newtonia confronts the evidence and arguments for special relativity, ought she to change her mind as to whether, as she puts it, \( F = \dot{p} \)? If you say yes and I say no, what is at issue between us? We agree on the physics of the matter, but disagree on what she means. I take her claim ‘\( F = \dot{p} \)’ to root her meanings, whereas you take her claim ‘\( F = m\dot{v} \)’ to do so. (We both agree that her ‘\( p = mv \)’ figures in rooting her meanings.) I thus take her to mean the same thing as we do by the sentence ‘\( F = \dot{p} \)’—namely that force is the rate of change of momentum, which is true. You take her to mean something different and false. You think she means something for which we normally have no words, though we can learn her language. What are we disputing, then, in this dispute over what she means? Perhaps, at root, we differ on the oughts of the matter, the empirical conditions under which she ought to accept or reject the various sentences she formulates.

Holding out this possibility raises many puzzles, but the puzzles may not be insuperable. I’ll mention two: In the first place, the meanings of a person’s words and

\[ ^{23} \text{I discuss this sense in Phil.Issues Huatulco piece.} \]
the contents of her thoughts have causal effects in the world. They explain things she does—whereas, we might worry, oughts don’t explain non-normative happenings. Horwich raises this as an objection to the claim that meaning is a normative concept. Normative explanations of non-normative happenings, though, are familiar enough in other realms: ethical theorists, for instance, debate the workings of “moral explanations”. How could the badness of a person’s character explain why people dislike him? Explain this, and we might have a template for explaining why, if the concept meaning is a normative concept, a person’s meanings can explain things she does.

In the second place, to the degree that Horwich meanings are determinate, they do seem to tie down what words mean. If there is a normative concept of meaning, how does it fit in with Horwich’s account? Why will the meaning of a word in a normative sense of ‘meaning’ be eligible as a meaning in Horwich’s non-normative sense? To this I would venture an answer with the same transcendental flavor as with analytic truths. The BAP of a word of mine is a set of ideal propensities I have. I’ll regard my ideal propensities as justified, barring any reason I can find not to do so. I’ll regard the things I have ideal propensities to do as things I ought to do. Take modus ponens, for instance: Since it accords with the BAP of my word ‘if’, I’ll not only accept inferences by modus ponens, but think that I ought to. A word means if, I’ll now think, just in case it has this normative quality, and I can take the normative quality to define meaning if. Meaning in this normative sense will thus, as I judge it, coincide with meaning, when determinate, in Horwich’s non-normative sense.

I hope I have conveyed, at least to some degree, how rich and original Horwich’s work on meaning is. Though it has many ties to current strands in philosophical approaches to meaning, Horwich gives novel twists to familiar elements in his theory. He comes up with an account that solves many of the problems that have confronted the theory of meaning from Quine to Kripke on Wittgenstein. My own thinking on meaning has been vastly transformed though grappling with his theory and arguments. Some of the worries I had when I worked on his book Meaning are addressed in Reflections on Meaning. I still have qualms, as I have indicated. The account of meaning is still rooted in an empirical conjecture whose truth would be stunning. Quine’s problem of indeterminacy remains, I have argued, in a more delimited form. And Horwich has not closed the door to thinking the concept of meaning to be a normative concept. But theory of meaning can’t be the same after this series of books by Horwich.
References


