Internalized Meaning Factualism

Jakob Hohwy

Abstract The normative character of meaning creates deep problems for the attempt to give a reductive explanation of the constitution of meaning. I identify and critically examine an increasingly popular Carnap-style position, which I call Internalized Meaning Factualism (versions of which I argue are defended by, e.g., Robert Brandom, Paul Horwich and Huw Price), that promises to solve the problems. According to this position, the problem of meaning can be solved by prohibiting an external perspective on meaning constituting properties. The idea is that if we stick to a perspective on meaning that is internal to meaning discourse, then we can preserve the normativity of meaning and yet locate meaning in the natural world. I develop a generic motivation for this position, but argue that, since this motivation is consistent with the Ramsey–Carnap–Lewis–Jackson reductionist strategy, internalized meaning factualism is unstable. The problems about the normativity of meaning can therefore not be sidestepped in this way.

Introduction

Meaning properties are normative properties, that is, there is a normative aspect to the fact that 'dog' has the property of meaning *dog*: 'dog' *should* be used of all and only dogs (*modulo* ambiguities, metaphors etc.). Many people worry that if we attempt to reduce the semantic to the non-semantic such as, for example, the physical, then the normative aspect of meaning cannot be retained. The intuitive problem is that from a set of physical, non-semantic, non-normative facts about, e.g., how someone has used 'dog' in the past, it is in principle impossible to read off how that person should use it in the future. Viewed at from a perspective *external* to meaning discourse, truths about such physical properties compel no unique interpretation of 'dog'. This clashes with the very natural intuition, from our

J. Hohwy (✉)
Department of Philosophy, Aarhus University, Jens Chr. Skous vej 7, Aarhus C 8000, Denmark
e-mail: filhohwy@hum.au.dk

perspective as normal language-users, that what we say *do* have a unique interpretation. The clash is strange enough to be called the Paradox of Meaning.[1]

If semantic properties cannot find a place in the physical world (I am using physicalism throughout for expository reasons, I could have focused on any non-semantic, non-normative ontology), then it seems we have to choose from a range of rather unattractive accounts of meaning: (a) *Meaning skepticism*: if there is no reduction of truths about meaning properties in terms of truths about physical properties, then it is wrong to say that there are meaning properties in the first place because the only *bona fide* properties we know are the ones we can locate in the physical world. (b) *Semantic primitivism*: there are irreducible non-naturalist semantic properties, perhaps of some Platonist sort, because we just know that there are normative meaning properties.[2] (c) *Elimination of normativity*: there are meaning properties, but once we know enough about the physical world we will be able to see that they are not after all normative. (d) *Semantic non-cognitivism*: claims about meaning properties are not in the fact-stating business in the first place. None of these four options seem particularly attractive: meaning skepticism seems as implausible, at least, as general epistemic skepticism; primitivism seems too easy, anyone can introduce Platonist entities to deal with all sorts of philosophical problems; eliminativism seems phenomenologically misguided; and non-cognitivism belies the fact that discourse about meaning seems at least minimally truth–apt.

A new kind of response is becoming increasingly popular, namely what I shall call *Internalized Meaning Factualism*. This response begins with a diagnosis of why the paradox of meaning arises: the culprit is what I above mentioned as the external perspective on meaning properties – the perspective which is external to discourse about meaning. If the diagnosis is right, then a cure would be to block the external perspective. Accordingly, the internalizers provide an argument to just this effect. Here I shall critically assess this ingenious new position in the debate about meaning.

I begin by identifying and setting out a generic version of the internalizing strategy, illustrating it with works by Paul Horwich, Robert Brandom and Huw Price. Then I discern how one might give a non-*ad hoc* motivation for this strategy, i.e., a motivation which is independent of worries about the paradox of meaning. Finally, I show how the strategy for reductive explanation associated with Ramsey–Carnap–Lewis–Jackson[3] is consistent with the internalizing strategy, and how the paradox reappears.

Internalized Factualism

The internalizers begin by considering this kind of argument: assume, for *reductio*, that a set of physical properties *f* constitutes *S*'s meaning *dog* by 'dog'. Now adopt a perspective of *f* external to discourse about meaning, that is, look at *f* as just a collection of physical properties, in particular look at *f* in isolation from any normatively laden truths about what *S* means by 'dog' or any other expression, or *S*'s mental states. From this external perspective, form hypotheses about the correct interpretation of 'dog' in *S*'s language. It seems that no matter how hard we were to look (and what our epistemic powers were), we

---

[1] Here I am trying to give the flavour of Kripke's influential arguments about rule-following, Kripke (1982).

[2] See, e.g., Boghossian (1989).

[3] See [Lewis (1970, 1972)] 1994; Jackson (1998).

would not find anything about *f* that could determine the normative aspects of meaning, i.e., that *S should* use 'dog' for the next Great Dane, rather than the next gothic cathedral or.... Truths about physical properties are normatively inert and therefore cannot decide between hypotheses with normative content. So *f* cannot be meaning-constituting. If this argument is right, then it bodes ill for the view that meaning could be factual in a naturalist sense.

Luckily, according to the internalizers, the argument isn't right because we can do something else besides looking at *f* from such an external perspective. We can *internalize* and *implement*: simply assume *f* is true of *S* (whether by this we mean that *S* has certain neuronal processes, or behavior dispositions, or structures that realize complex inferential roles for 'dog', etc.), and let *S* then use 'dog' to specify its own meaning (by saying, e.g., "'dog' means *dog*"). If *f* really does constitute the meaning of 'dog', then *S* will, in normal circumstances (i.e., sincere, not drugged etc.), be right. Importantly, from this internal perspective, where *S* is simply implementing the relevant meaning properties, questions of multiple interpretability do not arise.

What about us, the interpreters of *S*? Since what is internal to *S* is presumably external to us, don't we still have the problem of giving the right interpretation of *S*? No, because we simply also internalize and implement *f*, and thus we likewise put ourselves into a position where we can truthfully say of *S*: "*S* means *dog* by 'dog'".

The basic idea is rather commonsensical: if you want to know what something's proper function is (think of a shoe horn or a ruler, say), then learn how to use it yourself. Just looking at shoe horns and rulers will not get you anywhere.

The strategy is called internalized *factualism* because what drives the strategy is that there *are* facts about meaning that can make what *S* says true, when she says, e.g., "by 'dog' I mean *dog*". Without factualism, the internalizing strategy will be no help: there would be nothing to internalize and implement.

Different internalizers have very different philosophical agendas and sets of favored meaning-constituting properties (behavioral, dispositional, inferential role properties etc.). Common to them, though, are two concerns: first, that meaning-constituting properties cannot be non-naturalist, or Platonist, they must somehow fit in with the natural, non-normative world. They thus agree with most people, and in particular with reductionist physicalists, that the solution to the paradox of meaning should not be found by enriching our ontology with *sui generis* semantic properties.[4] Second, that if we conceive their favored meaning-constituting properties from a perspective outside actual discourse about meaning, then the normative aspect of those properties will be lost. These two concerns stand in tension because, straight off, nothing seems to prevent us from adopting an external perspective on natural, non-normative properties. Hence the internalizing strategy: embrace the notion that meaning-constituting properties fit in with the natural, non-normative world, but provide an argument that blocks the perspective external to meaning discourse. We may call it 'representation-by-doing'.

Paul Horwich employs an internalizing strategy. He argues that meaning properties, such as *S*'s meaning *dog* by 'dog', are constituted by basic acceptance properties. A basic acceptance property is a use property, that is, a certain regularity in use (e.g., certain of *S*'s dispositions, inclinations, or tendencies to use 'dog' for all and only things that are clearly

---

[4] Horwich's account is explicitly reductive, though, of course, in such a way that the paradox will not arise. Brandom explicitly distances his position from Platonism. Price's position is designed to avoid the paradox while also avoiding Platonist non-naturalism.

dogs), such that this regularity best explains S's overall use of 'dog'. Such regularities are non-semantic, non-normative, ultimately physical, properties. Horwich then employs the internalizing strategy in explaining how we get from the non-semantic to the semantic:

> [T]here will be no way to read off which meaning is constituted by a given use property. The best we can do, in order to get from one to the other, is to appreciate that some word (say, 'glub') has the use property – i.e., to actually use it in that way; in which case we can deploy that very word to characterize the constituted meaning (as "x means GLUB"). (Horwich, 1998, p. 66; Cf. also p. 89; 107–112).

Thus there is no external perspective from which we can read off the meaning of 'dog' from the physical use property which constitutes its meaning, we have to internalize that very property and discharge the explanatory task by simply using the word in question to explain what it means. From the internal perspective there will be no question of multiple interpretations: there are only particular uses of 'dog' which all can be explained by the basic use property which constitutes its meaning. We might think that this is not a particularly satisfying or substantial explanation, but that objection is not very strong: the core issue is not that we get a substantial, interesting explanation of meaning, but that we get an explanation which avoids the paradox of meaning. And here Horwich's approach seems promising.

I think we can discern a similar strategy in Robert Brandom's brand of inferential role semantics. According to Brandom, intentional content is a function of the commitments and entitlements associated with making moves in language games, that is, associated with inferential roles that can be represented in normative, but non-semantic terms. If this is going to issue in a factualism about meaning, then it had better be the case that when we interpret someone, then some interpretations are better than others. That is, when we obtain an external perspective on some stranger's linguistic behavior, it had better be the case that there isn't an indefinite number of equally good interpretations. Here is how Brandom employs the internalizing strategy in explaining how the problem of multiple interpretations doesn't arise:

> [T]he norms governing the use of the home idiom determine how to project the concepts used to specify the content of the stranger's attitudes (which determine how it would be proper to apply those same concepts in new situations) in the same way as they do for the ascriber's own remarks. [...] Thus the collapse of external into internal interpretation means that the problems caused by the gerrymandered alternatives to any particular discursive interpretation of another community from the outside is displaced to the context of interpretation and projection within our own community. This regress to our own interpretive practices dissolves, rather than solves, the gerrymandering problem concerning the relation between regularities and norms. For there is no general problem about how, from *within* a set of implicitly normative practices, what we do and how the world is can be understood to determine what it would be correct to say in various counterfactual situations – what we have committed ourselves to saying, whether we are in a position to get it right or not (Brandom, 1994, pp. 647–8).

Brandom supports this internalizing strategy by giving a detailed story about how our ascriptions of contents to interlocutors, whether they be from our own community or not, essentially involves engagement of our own norms (in order to make *de re* ascriptions), as well as assessment of the interlocutor's commitments (in order to make *de dicto* ascriptions) (Brandom, 1994, Ch. 8).

An internalizing strategy is explicitly employed by Huw Price (1997) in the context of an attempt to defend discourses about meaning, morality, modality and the mental from disappearing in a natural world. Price begins by endorsing Carnap's famous distinction between internal and external ontological questions:

> Carnap argues that there is no absolute, theory-independent ontological viewpoint available to metaphysics. Ontological questions about the entities mentioned in a particular theory or linguistic framework can properly be raised as what Carnap calls 'internal questions' – questions posed within the framework or theory in question – but not as 'external questions', posed from a stance outside that framework (Price, 1997, p. 250).

The Carnap thesis can then be used to claim that meaning discourse can retain an adequate factual status because the paradox of meaning only would arise if we ask ontological questions about meaning properties from an illegitimate external viewpoint. If meaning discourse is factual (in contrast to, say, discourse about the comical), then, as long as we stick to the standards that govern this discourse, there will be no opportunity for the issue of multiple interpretations to arise. That issue only arises if we cease to conceive of meaning properties as normative, that is, if we move to another, say, physical, discourse.

One classical way that the problem of multiple interpretations arises is from the perspective of the radical translator who conceives of meaning-constituting properties as the natives' dispositions for verbal behavior. On Price's view, as I understand it, this perspective would be outlawed: once we view meaning properties as physical properties we have left the standards that govern meaning discourse behind, and thus we have adopted an external perspective on the properties in question. If we instead stick with the perspective of the natives themselves, then, because meaning discourse is factual, we can simply use the natives' words to state what their correct translation is by saying, for example, "'gavagai' means gavagai".

Notice that Price presents the external viewpoint as absolute, theory-independent and ontological. This formulation would seem to be irrelevant for the physicalist bent on reductive explanation, and irrelevant to the project of defending the volatile discourses mentioned above from the 'rise of science' (Price, 1997, p. 247). For physics (and science in general) does not adopt such a theory-independent viewpoint; rather, physics adopts a viewpoint that is dependent on physical theory. So if the Carnap thesis is going to be relevant for the reductionist issue at hand, then it should be read in a more modest way: ontological questions concerning certain properties can only be raised within the discourse whose canonical theory quantifies over the properties in question (Price does consider this version of the thesis, 1997, sec. V, VI).

I have left out much of the detail of Horwich, Brandom and Price's otherwise very different stories, but I think it is clear that they all employ something that is recognizably the internalizing strategy: allowing meaning factualism (in various guises), but avoiding the paradox of meaning simply by – in the context of ascription of propositional attitudes, interpretation, or inquiry about meaning – letting speakers or interpreters use the interpreted terms, in the context of discourse about meaning, to express what their correct interpretation is. Horwich cashes this out by demonstrating how getting to master a use regularity can enable a speaker to use the word in question to state its meaning; Brandom cashes it out by showing how interpretation relies on engaging the normative properties of the interpreter; Price by allowing that discourse about meaning can be unproblematically descriptive and hence allow its participants access to semantic properties.

Now we can set out a generic version of the internalizing strategy. The paradox of meaning is this: on the one hand, since we do engage in meaningful communication, there are truths about the correct use of linguistic expressions. On the other hand, since no truths about non-normative properties could put normative constraints on linguistic use, there are no truths about the correct use of linguistic expressions.

The internalizers embrace the first part of the paradox by emphasizing that, internal to meaning discourse, issues of multiple interpretability do not arise. And they block the argument that leads to the second part by arguing that it relies on a perspective that is illegitimate because it is external to meaning discourse. The embracing and the blocking moves are equally important. To be satisfied we have avoided the paradox we should be told how meaning can be constituted as well as how the problem of multiple interpretations can be prevented from arising.[5]

Motivating the Internalizing Strategy

The embracing move of the internalizers' strategy looks promising, but without the blocking move it will be worthless. And, as it stands, the second move looks less than convincing: it says 'if you want to avoid the paradox, then don't look at certain properties in certain ways'. It is like saying 'look, death is a natural phenomenon, and if we refrain from looking at it from a psychological perspective, then there is nothing at all to be sad about'. So we need to independently motivate the blocking move.

Horwich argues that the external perspective relies on inflationism about truth because from the external perspective we are compelled to look for substantial semantic relations. Brandom argues that the external perspective collapses into an internal perspective because in external interpretation one essentially engages in the kind of normative inferential practices characteristic of interpretation within a discourse. Price does not offer a motivation for the Carnap thesis except to say that it gives rise to a defense of the volatile discourses that science threatens; and he adds that the absolute theory-independent viewpoint is obviously unobtainable (though we have seen that this latter notion is irrelevant). Price however emphasizes that it is allegiance to the distinctive descriptive standards of various discourses that allows us to ask ontological questions within them (e.g., Price, 1997, p. 262) and this, I think, is to suggest that one cannot ask the relevant ontological questions in other discourses that are not governed by those standards.

We can discern the gist of an independent general motivation for the internalizing strategy in these remarks. The overarching story is that factual talk of meaning and interpretation is inevitably tied to the standards that govern discourse about meaning, as well as appropriate parts of the discourse to which the terms in question (e.g., 'dog') belong; therefore factual talk of meaning and interpretation cannot happen in discourses not

---

[5] Another internalizing strategy would be of a more idealist bend, see, e.g., [McDowell (1977, 1994)]. I do not discuss this here since it seems to me to get rid of the tension between physics and semantics by denying that there are any non-normative facts at all. We also find something like the internalizing strategy in Davidson's writings, e.g., Davidson (1990), where, e.g., interpretation is explained via the metaphor of triangulation which seems to exclude an external perspective. And Wittgenstein (1953), §201 suggests the same: that there is a way of following a rule which is not based on interpretation of a given rule.

governed by those very standards.[6] What we need to know is *why* talk of meaning and interpretation must be tied to those standards.

For this story to unfold we begin by telling it in the ontological mode favored by Price, so that it concerns questions like "does 'dog' have a meaning?" (or "is there a unique interpretation of 'dog' in L?"). We standardly answer such questions by moving within what we can call meaning discourse (M-discourse), that is, by addressing the following types of question: Can 'dog' be used in sentences that convey information about dogs? Do normal speakers in normal conditions normally call all and only dogs 'dog'? Is the use of 'dog' guided by its meaning? Would we say that if the use of 'dog' were to change systematically, then its meaning would change too? Can normal, competent speakers be mistaken in their application of 'dog'? Do normal speakers have some kind of privileged knowledge of the meaning of their word 'dog'? And so on (to be completed in accordance with what one thinks is most salient to meaning discourse). If we answer 'yes' to enough of these questions, then we would presumably say that 'dog' does in fact have a meaning (or that there is a unique interpretation of 'dog' in a particular language L).

Now say that being a *participant* in a discourse D is having the core beliefs, or mastering the standards of assertability, or the salient inferential moves, or the platitudes, that make D the discourse it is; for short, being a participant in D is mastering the D-theory (for example, being a participant in color discourse is mastering such platitudes as that 'red' applies to all and only surfaces that are clearly red, that red is more like orange than it is like yellow, and so on). Assume that the above questions concerning meaning and dogs – that allowed us to debate whether 'dog' has a meaning – represent the core beliefs of M-discourse, that is, that they represent M-theory.[7] Then we can say that being a participant in M-discourse is mastering M-theory.

It seems intuitively clear that for us to understand the ontological question about meaning or interpretation, and begin appreciating answers to it, we need to be participants, in the above sense, in M-discourse. Conversely, if one is a participant solely in, say, physical discourse, P, governed by P-theory, then one will not understand and appreciate the ontological debate about meaning or interpretation. An explanation of why this seems intuitively clear is that the meaning of a term t is constituted by the platitudes or core beliefs that govern its correct use, and those platitudes or core beliefs are precisely what can be represented as the D-theory for a given discourse D (e.g., M-theory for M-discourse). Thus it will be impossible to engage questions of meaning without engaging M-theory. At heart, this explanation is based on a functional role semantics: t's meaning is given by the network of input, output and internal relations characteristic of t. If we say that M-terms are the terms characteristic of M-theory, then we can now say that understanding and appreciation of M-terms depend on mastery of M-theory, or participation in M-discourse.

I said that at heart this is functional role semantics. But there is a crucial twist to the internalizers' take on the functionalist story. It is not just that the meaning of 'dog' depends

---

[6] It is perhaps difficult to see how Horwich's deflationism suggests this kind of reading. The thought is that it is this kind of internalist allegiance to standards which allows us to make sense of the right hand side of the deflationary schemas for 'refers' and 'true': $(y)[N$ refers to y iff $N=y]$ and $(y)[$predicative concept F is true of y iff $Fy]$. Contrast this to the standard-independent inflationary schemas: $(x)(y)[$singular concept x refers to y iff $Cxy]$ and $(x)(y)[x$ is true of y iff $Rxy]$, where 'C' and 'R' are some substantial (e.g., causal) relations (see Horwich, 1998, p. 108).

[7] Had we been discussing whether 'wallaby' has a meaning, then M would contain not the relevant segment of discourse about dogs, but an appropriate segment of discourse about medium sized marsupials.

on the characteristic functional role of that word. It is also that our very inquiry about meaning, and engagement in interpretation, is conducted in a discourse that likewise is associated with a characteristic set of functional roles, those represented by *M*-theory. This is the theory that allows us to tell what meaning properties are, and what the meaning properties of particular words may be – it is what allows us to represent facts about meaning. This is a central part of the internalizers' story because when *S* internalizes and implements a meaning property in order to use a word to *say* what its meaning is, then *S* is thereby participating in *M*-theory. Moreover, it is a crucial part of their overall strategy (discussed in Internalized Factualism above) because participating in *M*-theory is the 'representation-by-doing' characteristic of this type of positions.

Now we can begin to see the independent motivation for the internalizing strategy: one must oneself come to master *M*-theory and thus participate in *M* in order to conduct actual inquiries about meaning, or engage in actual interpretive practices. If not, then one simply cannot ask the right sorts of questions or appreciate the relevant sorts of answers. This is a way of saying that one simply cannot debate meaning or engage in interpretation from a perspective – or discourse – external to *M*: as soon as inquiry or interpretation is under way one must be becoming a participant of *M*. Let us state this as the Internalizers' Principle: You cannot inquire about meaning or engage in interpretation from a perspective external to meaning discourse because representation of the very facts about meaning depends on actual participation in meaning discourse. Notice that this is not an *ad hoc* motivation since it rests on a substantial functional role account of meaning: meaning depends on *D*-theory. One could have alternative motivations which did not rest on that account of meaning, and which would not block the external view (such as, e.g., a causal account of meaning).

Is it fair to say that roughly this kind of functionalism is what motivates our three internalizers' positions? I think so. Firstly, they are all use-theorists about meaning and as such it is hard to believe they wouldn't endorse a functional role semantics under *some* description. Secondly, this kind of functionalism seems apt to capture what it takes to discharge their specific implementations of the internalizing strategy. Horwich: we cannot read off some meaning-fact what a word's meaning is, we must use the word to characterize what its meaning is, that is, only by engaging *M*-theory (e.g., "'...' means __") can we characterize its meaning. Brandom: external interpretation collapses into internal interpretation because the first essentially engages the *M*-theory characteristic of the latter. Price: we cannot ask ontological questions about meaning if we violate the Carnap thesis because violating the thesis is tantamount to leaving *M*-theory behind.

The Internalizing Strategy and Reduction

We have now seen the principle concerning meaning that appears to underpin the internalizing strategy, viz. that interpretation and inquiry about meaning depend on participation in *M*-discourse. It is clear how this principle can be put to service in the internalizing strategy: what brings the normative inertness of the physical to the fore is the kind of reductive explanation that seems to require inquiry about meaning, or engagement in interpretation, from a perspective external to *M*. How else can we, as it were, stand in *P*-theory and yet attempt to talk about meaning? But this would violate the internalizers' principle.

This really relates to an old question: what to do if the proposed reducing theory does not contain the characteristic terms of the theory we are hoping to reduce? The internalizers tell us to rejoice! Their argument is that the *P*- and *M*-theories must be kept apart, and that as long as they are kept apart there will be no paradox.

For the sake of argument, I shall assume the internalizers' principle, and then examine whether blocking the external perspective will prevent the paradox of meaning from arising while retaining the internalizers' other aspirations. It transpires that the principle does not prevent reduction and that internalism therefore is destabilized.

We begin by re-telling the story about how meaning depends on a D-theory in terms of implicit definition. M-theory was the collection of characteristic standards, platitudes or core beliefs of M-discourse. The basic idea was that the meaning of an M-term f_i depends on M-theory, so, since M-theory of course itself contains f_i, that is just to say that M-theory implicitly defines f_i, We can then (having systematically substituted all the characteristic predicates in M-theory for property names) represent the idea that M-theory implicitly define M-terms like this:

$$\#f \tag{1}$$

where 'f' is a schematic name for all the terms characteristic of M, and '#' names the remaining part of M-theory, the part that does not contain any of those terms. Notice that since # is stripped of M-terms it contains only terms that we must presume are neutral in the sense that they are not distinctively semantic or normative. To say that $\#f$ is true is to say that 'f' names the properties that make # true.

So far we have been working with the materials supplied by internalized factualism. However, at this point the scene is in effect set for applying the Ramsey–Carnap–Lewis–Jackson program for reduction.[8] We begin by substituting all occurrences of the property names for variables bound by an existential quantifier, to get:

$$\exists x(\#x) \tag{2}$$

that says that something best satisfies the rest of M-theory. Now we combine (1) and (2) such that we are neutral as to whether there are in fact any fs:

$$\exists x(\#x) \rightarrow \#f \tag{3}$$

saying that if something best satisfies #, then it is f. Now we can turn the implicit definition of f into an explicit definition:

$$f = \iota y \forall x(\#x \leftrightarrow y = x) \tag{4}$$

that says that f is the thing, whichever it is, that best satisfies the remaining M-theory, #; (4) thus gives the functional role for f. This crystallizes what lies behind the internalizers' principle: meaning depends on mastering the platitudes or core beliefs of the appropriate discourse, and, in particular, the meanings of the characteristic terms of meaning discourse depend on mastering M-discourse.

Having adopted the internalizers' principle in the guise of (4), then, the crucial question is whether it can prevent the reduction of M-theory to P-theory.

One could expect reduction of M-theory to P-theory to require bridge laws that connect terms of one theory with the other. Finding such bridge laws would presumably be an empirical matter of finding substantial relations between two sets of properties named by those terms, and it would presumably require that one obtain an external perspective on the two theories in order to see how they 'fit'. It seems obvious that such a method of reduction would violate the internalizers' principle.

---

[8] [Lewis (1970, 1972); Jackson (1998)]; I follow Lewis (1972).

However, with the Ramsey–Carnap–Lewis–Jackson approach one does not need to discover such bridge laws in order to reduce M-theory to P-theory. Assume that we, wholly in the f^* terms of P-theory together with #-terms neutral with respect to M-theory and P-theory (and thus independently of any M-theorizing), find that:

$$f^* = \iota y \forall x (\#x \leftrightarrow y = x) \tag{5}$$

That is, assume that we find that f^* is the physical property that best satisfies the functional role given in the neutral terms on the right hand side of the first identity sign. Then, due to the similarity between that which defines the f-terms characteristic of M-theory and the f^*-terms characteristic of P-theory, we can *deduce*:

$$f^* = f \tag{6}$$

And this bridge principle gives us the reduction of M-theory to P-theory since it allows us to locate the semantic wholly within the physical.[9]

Now, have we in the foregoing violated the internalizers' principle that one cannot inquire about meaning or interpret from a perspective external to M-discourse? It would appear that we did not violate the principle: the principle is not relevant to (5) because (5) is couched solely in P-terms, and (6) simply follows from the story told in P-terms, together with (4). Therefore we are at no point required to participate solely in P-theory and yet attempt to inquire about meaning or engage in interpretation. So the internalizers' principle does not preclude the reduction of meaning to physics. But, since we hereby are able to represent that meaning properties are physical properties, we will have no reason to believe that the *paradox* of meaning has gone away because, as we discussed at the beginning, it is a presupposition of the whole discussion that physical properties are normatively inert – the entire internalist strategy is designed to keep these properties out of the story. Hence, even as motivated by the internalizers' principle, the internalizing strategy does not make the paradox of meaning go away. 'Representing-by-doing' is not going to allow the non-reductive naturalist to have their factualist cake and eat it too.

We can represent this in terms of the embracing and blocking moves described above. The idea was that those moves give us reasons to believe that there is no paradox of meaning because we are told how facts about meaning come into place when we embrace the internal perspective and we are told how the external perspective is blocked. But these reasons are defeated in the light of the Ramsey–Carnap–Lewis–Jackson reduction strategy. If $f = f^*$ is true, *and* we have reason to believe that f^* is normatively inert, then we have no reason to believe that the embracing move is going to work, irrespective of what we think about the blocking move.

So the paradox reappears: we have reason to believe that S can inquire about meaning and engage in interpretation, but we also have reason to believe that S can't. And, crucially, the paradox reappears without ever having to adopt an external perspective on M-discourse.

None of this is to say that the Ramsey–Carnap–Lewis–Jackson reduction of meaning to physics is going to be successful. To stress, it is an assumption (that fuels one half of the paradox) that the non-semantic physical facts are normatively inert and that the Ramsey–Carnap–Lewis–Jackson reduction therefore cannot work. The above argument shows that internalist factualism builds on functionalism and therefore invites the Ramsey–Carnap–

---

[9] Steps (1)–(6) follows Lewis, (1972). I have suppressed some of the issues concerning this type of reduction, e.g., whether there has to be a unique realiser (see also Lewis 1994).

Lewis–Jackson strategy, whereby it cannot avoid the paradox even when internalism is respected.

Of course, the internalists could accept that they are wedded to this type of reductionist strategy but insist that we should be neutral about whether the strategy is going to be successful or not. This would be slightly odd because the entire debate began with assuming that non-semantic properties are normatively inert but perhaps the interalizers could insist that on the approach they advocate, the question of the normative status of the physical doesn't arise so they cannot be said to presuppose anything about this issue. Even so, the internalizers cannot retain all their aspirations. Assume first that Ramsey–Carnap–Lewis reduction *is* successful. In that case *M*-theory captures the normative aspects of meaning, and truths about the physical properties that satisfies the theory, together with the remaining neutral terms of *M*-theory, will *a priori* entail those normative truths. But in that case we have reduction of meaning to physics, contra the core anti-reductionist aspirations of the internalizers. Internalism would then simply be redundant. This would be the response favored by reductionists. Assume instead that Ramsey–Carnap–Lewis–Jackson reduction turns out to be unsuccessful. In that case there is no physical satisfier (or in general, no non-semantic satisfier) of *M*-theory. This failure could be explained in two ways, giving us two different alternative conclusions. (a) *M*-theory is correct and meaning is something over and above the non-semantic since *M*-theory would be satisfied by *sui generis* meaning properties. The problem is that this makes it impossible to sustain a plausible non-Platonist naturalism, again contra the core aspirations of the internalizers. This would be the response favored by those who defend Platonist primitivism about meaning. Or, (b) there is no physical satisfier of *M*-theory because that theory doesn't capture the essential aspects of meaning; *M*-theory is therefore incorrect and irrelevant for representing meaning properties. In this case, we would need a better theory to see if a reduction is feasible. The implication is that the internalizers' main tool, *M*-theory, for explicating meaning simply is inadequate. This would be the conclusion favored by those skeptical of functional role semantics, or those who think that functional roles should be open for revision in the light of philosophy.

It is useful to compare with how the analogous debate plays out in philosophy of mind. There, *a priori* (Ramsey–Carnap–Lewis–Jackson reduction style) physicalism will turn out to be true if concepts of phenomenal properties are functionalizable and have physical realisers. *A priori* physicalism will be false if the concepts are not functionalizable or there are no physical realisers. What we have just seen is that the internalizers argue that concepts of semantic properties are functionalizable (in *M*-theory) and that they cannot say that there are no non-semantic realisers without losing their naturalism; this is what opens for the reduction. *A posteriori* physicalism in philosophy of mind, by contrast, will be true if there are (in principle) empirically discoverable necessary identities of the phenomenal and the physical, and false otherwise. The internalizers should not be seen as adopting this kind of position because it is incompatible with internalism: it requires that we can phrase meaningful clearly external questions such as "is *this* physical property the meaning-constituting property for 'dog'?". *A priori* and *a posteriori* physicalism both seem to be consistent positions, internalized factualism is not. In fact, I think the internalizers see themselves as outside this type of debate. They do not want to privilege any discourse, their naturalism is not to be confused with physicalism. Their naturalism is the view that each discourse (moral, mathematical, modal, meaning, physics etc) can be naturalist as judged by its internal standards. My argument is that they achieve this by committing to functionalization of the discourses and thereby to reduction.

Price in fact anticipates that the internalizers' strategy doesn't rule out reduction claims like (6):

> It might be said that [Price's pluralist version of the internalizers' strategy] doesn't show that the identity claim [(6)] is false, merely that it is unmotivated. However, is falsity what the pluralist requires? A better approach seems to be that the identity claims are simply not well-formed, unless they respect the lessons of the Carnap thesis. (Price, 1997, p. 265).

But, as I have argued, the identity claims do respect the lessons of the Carnap thesis: (6) follows without having to adopt an external perspective on matters of meaning. And, anyway, the claim that (6) is not well-formed sounds odd given that (6) follows deductively from (4) and (5). The claim that (6) is ill-formed is thus akin to the obviously false claim that, if a and c appear to be different in some sense, then the expression $a=c$ is ill-formed, and therefore cannot follow from the premises $a=b$ and $b=c$ (think for example of the Morning Star, the Evening Star, and Venus). It seems to me, therefore, that the claim that (6) is not well-formed must after all come down to something like the claim that it is not motivated. But this seems to be a weak response. Firstly, it might be that a motive for reduction can be found (ask a couple of cognitive neuroscientists, perhaps). Secondly, given the right sort of physical story the semantic story follows deductively, so saying that the inference is unmotivated is a bit like saying that $a=c$ doesn't follow from $a=b$ and $b=c$, as long as there is no motive for performing the inference; and that is obviously wrong since the truth doesn't go away for want of motive to look into things.

The internalizers might be right that to ensure that we, from the outset, are talking about meaning properties and not something else, we must stay within M-theory. That is how we define our topic. But it doesn't follow that those meaning properties cannot then be located somewhere else, perhaps among the physical properties.[10]

References

Boghossian, P. (1989). The rule-following considerations. *Mind*, *98*, 507–550.

Brandom, R. (1994). *Making it Explicit*. Cambridge, MA: Harvard University Press.

Davidson, D. (1990). Three varieties of knowledge. In A. Phillips Griffiths (Ed.), *A. J. Ayer; Memorial Essays, Royal Institute of Philosophy Suppl.* (p. 30). Cambridge: Cambridge University Press.

Horwich, P. (1998). *Meaning*. Oxford: Oxford University Press.

Jackson, F. (1998). *From metaphysics to ethics*. Oxford: Clarendon.

Kripke, S. (1982). *Wittgenstein on rules and private language*. Oxford: Oxford University Press.

Lewis, D. (1970). How to define theoretical terms. *Journal of Philosophy, 62,* 427–466.

Lewis, D. (1972). Psychological and theoretical identifications. *Australasian Journal of Philosophy, 50*, 249–258.

Lewis, D. (1994). Reduction of mind. In S. Guttenplan (Ed.), *A companion to the philosophy of mind*. Oxford: Blackwell.

McDowell, J. (1977). On the sense and reference of a proper name. *Mind*, *86*, 159–185.

McDowell, J. (1994). *Mind and World*. Oxford: Oxford University Press.

Price, H. (1997). Naturalism and the fate of the M-Worlds. *Proceedings of the Aristotelian Society, Suppl. Vol. LXXI.* 247–267.

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Oxford University Press.

---

[10] For a systematic account of these notions of defining the topic and locating properties, see Jackson (1998).