

## **What Are You Thinking?**

### **Character and Content in the Language of Thought**

Louise M. Antony

The Ohio State University

Here's a view of the mind that I find attractive. A mind is a kind of naturally-occurring computer -- an information-processing device. It comes with its own medium of computation, a system of symbols that constitutes, literally, a language of thought. Cognitive states, and perhaps mental states generally, involve functional relations to token symbols in this system. To call the elements of this system "symbols" is to impute to them two important kinds of property: lexical-syntactical, and semantic. That is, the symbols that constitute Mentalese have both form and content. This view of mind – familiar as the "Computational/Representational Theory of Mind" (C/RTM) – dovetails nicely with (although it does not entail) a view of natural language that I also find attractive. On this view, natural languages are overt, (partly)<sup>1</sup> conventional symbol systems. They are ontologically posterior to Mentalese, in the following ways: first, you need a mind in order to speak a natural language, but you don't need to speak a natural language in order to have a mind; second, the semantic or representational properties of the terms in a natural language derive from the semantic properties of Mentalese. If

---

<sup>1</sup> Specifically, what is conventional is the relation between any particular word and its extension. Words in a natural language are not natural signs of the things they stand for, as is true in many other animal communication systems.

we add this view of language to the C/RTM, we get what Putnam calls, disparagingly, “The Cryptographer Model of the Mind.” He gives what seems to me to be a perfectly fair of the view in Representation and Reality:

The mind thinks its thoughts in ‘Mentalese,’ codes them in the local natural language, and then transmits them (say, by speaking them out loud) to the hearer. The hearer has a Cryptographer in his head too, of course, who thereupon proceeds to decode the ‘message.’ In this picture, natural language, far from being essential to thought, is merely a vehicle for the communication of thought. (R&R, 10-11)

What I want to argue in this paper is that the CMM harbors underutilized resources for understanding the nature of intentional content and its role in psychology. Specifically, by taking seriously the idea that creatures who speak a natural language are operating with two separate, full-fledged linguistic systems, we can get some purchase on the long-running debate about the “width” of psychology. Content, I contend, is always wide – the only kind of content there is, is extension. This does not mean that I advocate a wide taxonomy for psychology, however. I want also to endorse the view that psychological taxonomies are essentially narrow. The aim of this paper is to explore the prospects – and acknowledge some of the problems for -- a theory of mind with these features.

I’ll start with a brief, tendentious history of the internalism/externalism debate.

First there was Frege, who posed a puzzle about identity statements: how can the cognitive significance of a true identity claim, like “the evening star is the morning star” differ from that of a truism, like “the evening star is the evening star”? If identity

claims are claims about things in the world, if the evening star really is the morning star, then it appears that the first identity statement could be saying no more than that the evening star is self-identical, and yet it clearly is saying something more, since not everyone knows that the first statement is true, while the second is knowable a priori. Frege considered, but rejected the possibility that identity claims are not about things in the world; that they concern, rather, the signs we use to refer to things in the world. In that case, the informative identity claim would say, “the term ‘the evening star’ refers to the same object as the term ‘the morning star.’” This would explain how the two identity claims could differ in cognitive significance, but Frege rejects the suggestion on the grounds that it shifts the subject matter of the original statement:

a sentence like “ $a = b$ ” would no longer refer to a matter of fact but rather to our manner of designation; no genuine knowledge would be expressed by it. But this is just what we want to express in a many cases. [\[Frege in Martinich, p. 186\]](#)

So the puzzle is, how do we get identity claims to remain object-language claims, rather than meta-language claims, and still account for a difference in cognitive significance that seems traceable to a mere difference in the signs used?

Frege’s solution was to insist that there must be something about the signs other than their mere difference as signs that did the trick:

if the sign “a” differs from the sign “b” only as an object (here by its shape) but not by its rôle as a sign, that is to say, not in the manner in which it designates anything, then the cognitive significance of “ $a = a$ ” would be essentially the same as that of “ $a = b$ ”, if “ $a = b$ ” is true. A difference could arise only if the difference of the signs corresponds to a difference in the way in which the designated

objects are given. [\[ibid.\]](#)

This “way ... the object is given” Frege dubs the “sense” of the expression; the object “given” by a particular sense is the “referent” or “nominatum.” Since the same object may be presented to a thinker in a variety of different ways – via many different senses -- and because the cognitive significance of a statement is determined by senses, rather than by referents, we have the desired account of the differences in cognitive significance between informative and truistic identity statements.

But what exactly is a sense? Frege told us some things about the properties senses must have in order to play their appointed rôle, but left unanswered many questions about their character. Thus he told us, as noted above, that the relation between sense and reference is functional, but that the relation between reference and sense is one-many. He also told us that the content of a thought is a function of the senses of the expressions that compose the sentence that expresses the thought, while the truth-value of the thought is a function of the referents of those expressions. Finally, he told us that in certain contexts, for example in “that-“ clauses, there is a systematic shift in the reference of terms – they refer, in such contexts, to what is customarily their own senses. Beyond that, we are told only that the sense is something that “lies in between” the object designated and the subjective images speakers may associate with a term. The latter may well vary from person to person, whereas the sense of an expression is or can be the “common property of many.”

Kripke made the influential suggestion that Frege’s “sense” could be identified with something like a definite description – an internalized specification of properties that determined a unique referent. The textual basis for this suggestion is a remark

Frege makes at the beginning of his discussion of singular expressions. He says: “The sense of a proper name is grasped by everyone who knows the language or the totality of designations of which the proper name is a part.” [187](#) And then there is a footnote: “In the case of genuinely proper names like ‘Aristotle’ [as opposed to complex designating expressions] opinions as regards their sense may diverge. As such may, e.g., be suggested: Plato’s disciple and the teacher of Alexander the Great. Whoever accepts this sense will interpret the meaning of the statement ‘Aristotle was born in Stagira’ differently from one who interpreted the sense of ‘Aristotle’ as the Stagirite teacher of Alexander the Great.” [197](#) According to Kripke’s reading (although of course Kripke was not primarily interested in Frege exegesis), then, Frege’s view could be assimilated to Russell’s, who explicitly argued that ordinary proper names should be treated as abbreviations for definite descriptions. The resulting “Frege-Russell” view of proper names became the object of Kripke’s critique.

What I want to focus on is the transformation of Frege’s sketchily characterized “sense” into something that the speaker believes about the referent of the term in question.<sup>2</sup> That is, the grasping of a term’s sense, on the view Kripke constructs, is a kind of propositional attitude: understanding the term “Aristotle” means believing that there is one and only one object with a certain set of properties. Here’s how Kripke puts the central clause of the view he wants to attack:

(1) To every name or designating expression “X,” there corresponds a cluster of properties, namely the family of those properties  $\phi$  such that A believes “ $\phi X$ .”

[\[N&N in Martinich, p. 259\]](#)

---

<sup>2</sup> Two philosophers who resisted this interpretation are Gareth Evans [\[Varieties of Reference\]](#) and Michael Devitt [\[ref\]](#)

Subsequent clauses tell more about the family of properties: that they are believed by A to be uniquely true of some object, that they in fact determine the referent of “X,” and that the speaker’s belief “If X exists, then X has most of the  $\phi$ ’s” is known a priori by A, and is a necessary truth in A’s idiolect.

Kripke’s argument against this view consists in a series of counterexamples designed to show that, for typical proper names, there is no set of properties meeting these conditions. The upshot, according to Kripke, is that proper names do not have senses – they have, as Kripke put it, echoing John Stuart Mill, denotations but no connotations. The positive part of Kripke’s view was his story about how denotations were fixed, if they were not fixed by something the user of the term believes: reference is determined, Kripke said, by a historical chain of usage, stretching from an initial “baptism” in which the term is stipulated to apply to a certain object, and up to the present day. The links in the chain are forged by users’ intentions to pass on the term, with its referent, to other speakers, who must, for their parts, have a coordinate intention to use the term to refer to whatever it is that the term has been referring to so far.

Putnam was thinking along similar lines with respect to “common names” – at least with respect to those common names that denoted natural kinds. In “The Meaning of ‘Meaning,’” he argued that nothing that a speaker associates with a term like “water” or “gold”, and nothing the speaker believes about the things the terms denote, suffices to fix the reference of such terms. Rather, a term’s denotation is determined, as in Kripke’s story, by an initial reference-fixing stipulation, followed by a chain of usage forged by speakers’ intentions to use the acquired term in the same way

it was used by the speakers from whom the term was acquired. This initial stipulation involved, just as it did on Kripke's story about "baptisms," an essentially indexical element – a reference to a substance (in the case of natural kind terms) picked out ostensively, and not by means of any of its properties. In both cases, the thought was that the baptizer could be massively mistaken, at the time of the naming, about the properties the thing, or kind of thing, actually possessed. (A side note: in this respect, the Kripke/Putnam theory of reference determination was neatly consonant the prevailing Quinean views about confirmation and analyticity. Fregeanism, as Kripke and Putnam construed it, entailed the existence of unrevisable beliefs, unrevisable because they were constitutive of the meanings of terms.)

Some of Putnam's arguments paralleled Kripke's. He argued, for example, that it was perfectly possible that gold could fail to be yellow, and yet, for all that, still be gold, an argument analogous to Kripke's argument that Aristotle might have failed to be Plato's student and yet still have been Aristotle. But Putnam also offered a different sort of argument: he constructed a thought experiment designed to show not only that speakers' beliefs did not fix reference, but that nothing intrinsic to a speaker could fix reference. The thought experiment is by now so familiar as not to require rehearsal. Suffice to say that Putnam invited you to share his intuition that there could be two molecularly identical individuals whose words differed in extension: my word "water" refers to the substance H<sub>2</sub>O, the kind of stuff that lies at the end of my referential chain, while my molecularly identical twin's word "water" refers to a chemically different substance, XYZ, the stuff that lies at the end of her chain. It is this set of different extrinsic properties that accounts for the difference in meaning between my term and

hers.

The moral, according to Putnam, was that we would have to give up at least one of the features of Frege's picture of the relation between sense and reference: we could give up the idea that something graspable, something intrinsic to the speaker, determined reference, or we could retain the idea that sense simply was that which determined reference, and give up the idea that senses were in the speaker's head. Putnam seemed partial to the second option -- "meanings ain't in the head," he liked to say -- but he did allow that one could conceptualize meaning as a vector, with both an internal and an external component: "narrow" content and "wide" content, respectively. Narrow content consists of the properties, beliefs, and images a speaker centrally associates with a given term: in the case of the term "tiger" it might include the belief that tigers are large cats, as well as a mental image of what the speaker imagines to be a typical tiger. It is, however, the external component that does the heavy semantic lifting. The wide content of "tiger" is the natural kind tiger, or the nature of tigers themselves, and it is this that determines the extension of the term. Anything I say about tigers using that word depends for its truth or falsity on the actual properties of those animals, regardless of whether my beliefs are correct or my images are accurate.

That doesn't mean that narrow content is idle, however. Putnam allowed that the internal component might still have a role to play in explaining my use of a term: my stereotype can help account for my tendency to apply the term to the things that I in fact apply it to. It's because I believe that tigers are big cats with stripes that I manage to call tigers "tigers" and manage not to call lions "tigers." And because stereotypes are not determinative of reference, they may also account for various mistakes I make.

They might, for example, explain my failing to call something (I don't know – a small albino tiger) a “tiger” even if it is one.

This suggestion of Putnam's led to the joining of issues in semantics with issues in the philosophy of psychology. For many reasons, it seems reasonable to think that, if psychology is going to generalize over thinkers in virtue of the contents of their thoughts, that the content in question is going to have to be narrow. This is because of (a) Fodorian reasons: science taxonomizes in terms of causal powers, and causal powers are individualistic, (b) thought contents are introspectible, and narrow content would seem to be introspectively available to thinkers, whereas wide content is not, (c) ordinary folk psychological ascriptions and explanations are largely de dicto, rather than de re, and that would seem to support the idea that such ascriptions and explanations are doing business with narrow, rather than wide contents. That's not to say that there's consensus on this point – by no means. Many philosophers contend that ordinary psychological explanation is, in fact, typically wide (e.g., Burge, Peacocke). Still others say that scientific psychology is narrow, but that folk psychology is wide, and so there is no possibility of lining them up. This brings us more or less to the present. I want to lay out the options here a bit more systematically, but first I want to comment on what I see as two serious missteps taken during the development of the issue, missteps that, once recognized, will mandate a reconfiguring of the narrow content debate.

The first misstep I've already called attention to – it occurred with Kripke's suggestion that a Fregean sense could be identified with a description. The second one involves Putnam's move from saying that nothing a speaker believes about the referent can determine a term's extension, to saying that nothing in the speaker's head

determines a term's extension. Let me comment on each in turn.

Definite descriptions – descriptions of any kind – seem to me to be particularly bad candidates for senses, the notorious footnote notwithstanding. To say that the sense of “Aristotle” is a property, like being the most famous student of Plato, could mean either of two things: it could mean that “Aristotle” is an abbreviation for “the most famous student of Plato” or it could mean that I use the term “Aristotle” in a way that ties it essentially to the property of being the most famous student of Plato – that I take it refer to whatever object has that property. Consider the first possibility: notice that we haven't left the realm of the linguistic – we're replacing a term with other terms. This is very like Russell's view, but Russell was clear that his proposal about ordinary proper names was not like Frege's proposal. Russell did not believe that reference was mediated by anything, at least not at the level of logically proper names. Ordinary names, on his view, did not refer directly, but this was because they were abbreviations for other forms of representation – definite descriptions. Contextual definition of ordinary names was meant to be part of a regimentation of ordinary language into a logically superior language, one in which it would be revealed that there was nothing to meaning but reference. So the suggestion that the sense of “Aristotle” is another expression seems not at all in the spirit of Frege's proposal.

But what about the other option, which is, after all, the one that Kripke explicitly pursues? Frege was not offering, nor was he even presuming, a particular theory of mind when he posited a sense as something that lay “between” the object and the subject, so it's hard to say what he might have thought about the idea that a sense is a belief or set of beliefs that the speaker holds about an object. Standard translations

suggest that he identified what we call “propositions” with “thoughts”, so perhaps the suggestion that senses of terms were components of thoughts is a reasonable one. There is, I think, a deep problem connected with this suggestion, one that I’ll describe below, in connection with my discussion of Putnam. For now, let’s focus on this question: which components of thoughts are we talking about? If Frege thought that the sense of “Aristotle” was the property of being the Stagerite teacher of Alexander, what then would he have taken to be the sense of the expression “the Stagerite teacher of Alexander”? (I’m willing to bet that what’s worrying me here is also what Russell was going on about in the “Grey’s Elegy” passage of “On Denoting.”) Clearly the two expressions, “Alexander” and “the Stagerite teacher of Alexander” must differ in sense; otherwise, we’ve made no progress on the problem of meaningful identities.

Furthermore, circularity threatens: Suppose that the property of being the Stagerite teacher of Alexander constitutes the sense of the term “Aristotle.” How do I, the thinker, put myself into epistemic contact with that property? If the way to insert that property into our beliefs is by means of its standard expression in English, viz., “the Stagerite teacher of Alexander,” then it seems we would need to grasp the sense of that expression before we can grasp the sense of “Aristotle.” But then what is the sense of “Alexander”? It better not turn out that the only thing I believe about Alexander is that he was the student of Aristotle, or we’ll just never make any headway toward epistemic contact with the gentlemen themselves.

So – alternative suggestion: the property associated with the term “Aristotle” the one that constitutes the sense of “Aristotle” is, simply, the property of being Aristotle. Note that this suggestion does not run afoul of Frege’s objections to the metalinguistic

analysis of informative identities: [need to think about proposal in connection with Kripke's Condition C, but wouldn't matter if it violated it] the proposal is not that grasping a sentence containing the term "Aristotle" involves thinking about the term "Aristotle." But there is another problem: consider the informative identity claim, "Hesperus is Phosphorus." On the current proposal, the sense of "Hesperus" is the property of being Hesperus, and the sense of "Phosphorus" is the property of being Phosphorus, but arguably, those "two" properties are the same. If the property of being Hesperus is a real property at all, it's one that is held necessarily by, and necessarily only by, Hesperus. The same is true for the property of being Phosphorus. But since Hesperus is Phosphorus, these properties have the same extension in all possible worlds, which makes them, at least on many theories of property identity, the same property.

But even if we adopt a hyper-intensional view of properties, so that, somehow, the two properties, being Hesperus and being Phosphorus, come out to be distinct, there's a question we've yet to confront. What is it for a speaker to grasp these senses? Again, as I said, Frege did not offer us a theory of mind – he didn't tell us what "grasping" was. All we know is that different senses induce different cognitive states in a competent speaker: to grasp the sense of "Hesperus" is somehow different from grasping the sense of "Phosphorus." So whether or not these expressions have properties as their senses, we still need an account of what it would be for these distinct properties to be grasped.

Well, I suggest, the answer has been staring us in the face: it's provided by the CMM. The sense of "Hesperus" is simply the mental code name for Hesperus – a

mental representation of Hesperus. The sense of “Phosphorus” is also a mental representation of Hesperus, but a lexically distinct one. The sense of a term in natural language is its coded translation in Mentalese. Given that, and given the fact that CMM individuates cognitive state types partly on the basis of the particular Mentalese sentences the states involve, we get the desired result, namely, that I can entertain, Hesperus thoughts without necessarily entertaining Phosphorus thoughts. The state of believing that Hesperus is Hesperus comes out to be as different from the state of believing that Hesperus is Phosphorus as it is from the state of believing that Hesperus is made of green cheese. Now I emphasize that I’m not making the exegetical claim that this is what Frege had in mind – I’m saying only that the proposal fills the bill. A Mentalese expression is, clearly a “mode of presentation” – it is a way that a speaker thinks of a thing. Treating senses as Mentalese expressions gives us all the advantages of the metalinguistic analysis with none of the drawbacks. In particular, the proposal gives us ultra-fine-grained psychological states – just as if we were thinking about the expressions in natural language – without the shift in subject matter.

This point is worth lingering on – I’m suggesting that to grasp the sentence “Hesperus is Phosphorus” is to put myself into a functional relation to some particular Mentalese sentence. I can’t tell you which one; that’s because I don’t speak Mentalese, I only think it. But according to the proposal, the Mentalese sentence will have the following feature: it will involve two lexically distinct terms, each of which refers to Hesperus. (The terms are distinct from each other in just the way that terms in natural language are distinct from each other.) When I token such a sentence in thought, I am not thinking about the terms that constitute the sentence, I am thinking about the thing

that the terms refer to. I am not, as it were, mentioning the Mentalese terms in thought; I'm using them.

The key here is that sense is not being construed as the embodiment or representation of information about the referent. As I've already argued, there is no point to treating sense in this way. Instead, sense is nothing more than a way of representing the referent, just as it should be. Also, notice that sense is not being identified with the meaning of the natural language term. On this proposal, there is only one component to meaning, namely reference. Treating sense as a new linguistic item in a special language gives us the effect of a second semantic factor without our having to answer knotty questions about what that factor is. This sort of second factor is, moreover, immune to Kripke's objections. My semantic competence with respect to the proper name "Hesperus" does not consist in my being able to cite properties that uniquely identify Hesperus. Nor does my ability to distinguish mental states involving my "Hesperus" representation from those involving my "Phosphorus" representation require my associating different properties with "Hesperus" than I associate with "Phosphorus". I don't really need to associate any particular properties with either name, just as Kripke's account of proper names would have it.

But now there's the question of the functional relationship between my Mentalese symbol for Hesperus, which I'm urging us to regard as the sense of the English expression "Hesperus," and the referent of that expression. Does my Mentalese expression determine the referent of "Hesperus"? The answer to this question, I have to say, is "no." But, I contend, the spirit of Frege's proposal about the determination of reference by sense will be respected. In order to explain how, I need

to say more about what I identified as the second misstep in the development of the notion of narrow content: the move from “none of a speaker’s beliefs fix the referent of a term” to “nothing in the speaker’s head fixes the referent of a term.” Here’s why I think this is a misstep.

Let’s begin with the picture Putnam was attacking. Like Kripke, Putnam assumes that a Fregean sense is a body of information about the referent of a term, internalized by the speaker, and introspectively available to the speaker. So part of the sense of “water,” on this picture, is the speaker’s beliefs that water is a liquid at room temperature, that it quenches thirst, that it fills the lakes and rivers, and so forth. But now we can ask the same questions we posed to Kripke’s Fregean. Are we to understand this to be the claim that the term “water” abbreviates “the stuff that is liquid at room temperature, quenches thirst, etc.”? If so, we seem to have a proposal about analytical connections between terms in a natural language, rather than a thesis about what mediates the relation between the referent (the object) and the subject.

The alternative, as before, is to treat the sense as the properties we take to be distinctive of, or the beliefs we have concerning the stuff to which we take “water” to refer. But while there seems, on the surface, to be something common-sensical about such a suggestion, in fact, when we press harder, we see that it’s difficult to figure out exactly what the suggestion comes to. What is it for us to be in epistemic contact with those properties? Is a belief about, e.g., liquid, a mental state that transparently connects us to liquids? If so, why not such a state to connect us to water, straightoff? If the state of believing that there’s something that’s a liquid at room temperature is a representational state, then why would a such a state involving liquid be conceptually

prior to the one involving water? Why, that is, when we speak of “water,” must our connection to the stuff the term denotes be mediated by a mental state with intentional content about something else?

The problem I'm raising is very similar to one that I've always had with pure conceptual role semantics. Such theories propose that the semantic value of one concept, say, my concept HORSE, is fixed, not by that concept's relation to water, but rather by its relation to other concepts. The intuition behind this suggestion is that a speaker couldn't really have a concept of *horse* unless they have certain beliefs about horses – that they are animals, that they typically make a “neighing” sound, that they have four legs and hooves, and so forth. In other words, no one could have a concept that means *horse* unless that concept was inferentially connected to concepts that meant *animal*, *neighs*, *hooves*, etc. But such an intuition is only captured, it seems to me, if we take for granted that the concepts to which HORSE is presumed to be inferentially connected already have their contents fixed. Otherwise, there would be no way to specify the concepts to which inferential connections are particularly important. Going holistic at this point only makes matters worse: now we have a host of meaningless concepts, all of which have to be connected to each other for anything to mean anything. But which patterns of inter-conceptual connections induce which meanings on which symbols? Can we make sense of there being constraints of this sort on a concept's meaning *horse*? The original, motivating intuition seems to me to be long gone.

Things are supposed to improve if we have a “two-factor” conceptual role theory: maybe, but if they improve sufficiently, I suspect it's because the referential factor is

### doing all the work.

The problem, once more, is that senses are being thought of as an incorporation of information about the extension, rather than simply as a way of presenting the extension to the thinker. Once more, the better way is to think of the sense of the natural language expression as an element in the internal code: the sense of “water” is a Mentalese term that denotes water. But in that case, it is no longer so clear that meanings “ain’t in the head.” If there could be, as I’m insisting there is, something in my head that denotes water, then it appears that there is something internal to me that determines reference. This could be so even if we grant, as we no doubt should, that my beliefs about water are insufficient to fix the reference “water.”

At this point, we need to get clearer about the various views in play in the debate about internalism and externalism. Let us, for the moment, consider only the issues regarding the content of thought, and leave aside the relation of those issues to the question of the determination of the semantic meaning of words. It’s a bit awkward to do so, since the original issues were framed in the context of a debate about natural-language semantics, but I think it’s still preferable to consider thought on its own. (Actually, it’s the absence of any comment on the potential difference between thought-contents and word-contents that may be at the root of the problems I was raising earlier about the intelligibility of the position Kripke and Putnam attribute to the Fregean.)

Externalism is standardly taken to be the view that thought contents are determined by, or are partly determined by, something in the world. Now this seems to be to be ambiguous between saying: (a) content is essentially relational: the content of a concept is that which the concept stands for, and what concepts stand for are

(typically) things outside the thinker's head;<sup>3</sup> and (b) what determines what a concept stands for is something external to the speaker. Claim (a) is the claim that thought is essentially "world-involving" – that I cannot have a contentful thought but that there is something in the world that that thought is about. The most extreme version of internalism denies (a), but there it's possible to accept (a) and still be internalist in virtue of denying, or qualifying (b). So there is really a range of options here:

Strong (or Solipsistic) Internalism: Concepts can be contentful even if nothing other than the thinker exists. Concepts have no "satisfaction conditions" – they have their contents independently of anything external to the speaker.

Weak Internalism: The satisfaction conditions for a concept are completely determined by what's in the head. The world contributes to the determination of content by contributing the things that satisfy the satisfaction conditions.

Weak Externalism: There are two independent factors that jointly determine satisfaction conditions: one in the speaker's head, and one in the world. The world contributes twice, once by contributing to the determination of satisfaction conditions, and once by supplying the things that satisfy those conditions, once determined.

Strong Externalism: There is no way to "factor out" an internal contribution to the determination of satisfaction conditions.<sup>4</sup>

Now I take it that the standard way of interpreting the Kripke/Putnam argument is as an

---

<sup>3</sup> It's certainly possible to have concepts of inner states or objects – I have several. I doubt that any externalist would deny this. So we have to define all versions of externalism in a way that allows that the thing in the world a concept denotes might be a thing in the speaker's head. I'll continue to speak of "things outside the speaker's head" with this qualification understood.

<sup>4</sup> Thanks to Joe Levine for helping me develop this taxonomy.

argument against weak internalism and in favor of weak externalism (or, perhaps in Kripke's case, of strong externalism, since he never explicitly endorses the idea that there's anything analogous to a stereotype associated with proper names). That is, the "Fregean" they attack holds that the thinker's beliefs determine a set of conditions, and that anything in the world that happens to satisfy those conditions is the content of the concept. But as I've been at pains to point out, it's problematic to hold that the content of one concept is asymmetrically fixed by the content of other concepts. This doesn't mean, however, that one cannot hold that something intrinsic to the thinker fixes the satisfaction conditions of the thinker's concepts.

Consider Fodor's "asymmetric dependency" account of the determination of intentional content. According to Fodor, my concept HORSE has the content *horse* just in case a certain pattern of nomic connections holds between horses and HORSE-tokenings.<sup>5</sup> This theory does involve the world in content-determination. The content your concept has is going to depend on what, in your world, lies at the other end of the nomic relations. But content-determination is also world-independent in the following sense: thinkers whose heads are the same – in the relevant respects – will always be thinking thoughts with the same contents. If Fodor's theory is right, then there is a very important sense in which meaning is "in the head."

Putnam and Kripke are, obviously, externalists of one sort or another. One might argue that they are both strong externalists, since, on both of their accounts, whatever material there is in the speaker's head plays no role at all in content determination. But the matter is not so simple. Consider Kripke's sketched account of the way reference is

---

<sup>5</sup> This is not precise. But don't make me go through it.

fixed and passed on from speaker to speaker (pardon the lapse back into talk of the semantics of natural language terms): speakers' mental states, and in particular, their intentions are crucial to the account. The initial namegiver must, first of all, have a way of picking out in thought the object they intend to name, and then they must intend to name it.<sup>6</sup> Subsequent users must, when they begin to use the term, intend to use it in the same way as does the person who is their source for the term. As Kripke notes, if I hear of someone called "Iphegenia" or whatever the example is and decide that that would be a nice name for my dog, I've essentially created a new name. I've not picked up the reference to Iphegenia. The story Putnam tells for natural kind terms is the same in its essentials, except that the relevant intentions involve the notion of "same stuff" or "same kind."

So we have, on the Kripke/Putnam story, "reference-fixing" intentions, as well as what we might call "reference-catching" intentions in play. These intentions may well involve descriptive content, since, as was noted above, it's necessary at some point that the intended referent or extension be isolated in thought, and they also involve – and this is something that Putnam emphasizes – an indexical element. In the paradigm case, reference is fixed by a demonstration: "I name this child 'Hermione,'" possibly with a pointing, in the presence of the child to be named.

Now we know about how demonstratives work; David Kaplan taught us. A term like "this" is an element that introduces a contextual element into our speech or our thought. Here's how. On Kaplan's account, words have a feature he calls "character" –

---

<sup>6</sup> Perhaps this is not strictly necessary; perhaps there are ways that a thing can acquire a name "passively", perhaps through people's coming to associate some sound or mark with that thing. But in that case, there will be a point at which someone must use the sound with the intention of picking out that thing by means of that sound. This and other related possibilities are irrelevant to the current point.

character is a function from context to content (reference, for our purposes). Some terms, like proper names, have a constant function as their character – “Aristotle” picks out the same individual, namely Aristotle, no matter who is speaking, or where, or when.<sup>7</sup> But other terms – indexicals like “I” “now” “here” and “this” – have characters that determine different contents depending on context. The character of each expression can be thought of as a rule directing us to the particular contextual parameter that is relevant to determining the content of that expression. The character of “I” is such as to deliver, as the content of the expression, the individual speaking, and so forth. The character of the demonstratives, “this,” “that” “these” and “those” is such as to generally require some kind of pointing or description to fix a content. (Sometimes the necessary descriptive material is explicitly expressed, and sometimes it’s merely implicated: “Hand me that” said in a situation in which it’s obvious I need a hammer, for example.)

Kaplan argues that there is an important difference between the role played by definite descriptions and that played by demonstrations in determining the content of a sentence. Definite descriptions, like “the first human being to step foot on the moon” contribute properties to the content of the sentence; they produce general propositions. Demonstrations, like “that guy there on the moon” contribute individuals – they produce singular, or object-dependent propositions. The sentence “the first human being to step foot on the moon will be famous” OMIT????

[go back and fix whether talking about words or concepts – maybe admit you’re going to be sloppy about it]

---

<sup>7</sup> Yes, there is more than one thing named “Aristotle.” And yes, it probably didn’t refer to Aristotle before he was born. Let’s not be tedious.

I now propose to make heavy use of the character/content distinction. First off, a very straightforward application of Kaplan's theory to the case of a reference-fixing demonstration will enable us to get a bit of a handle on the way weak externalism works. Imagine our namer, setting out to name a child: she points to the child – the only child around, let's assume for convenience's sake – and says, "This child shall be called 'Hermione.'" The content of this speech act (which happens to be a stipulation) depends crucially on the individual that is determined to be the content of the demonstrative expression, "this child:" it will be just that child who has been given the name "Hermione". Here's how the demonstrative element contributes to that determination. The character of "this X" directs us to a particular contextual parameter: to (something like) the most salient X in the speaker's immediate environs. Thus, the character of "this" is such as to introduce a world-dependent component into the determination of the larger expression; it contrasts, in this respect, with "child," which determines satisfaction-conditions independently of features of context.<sup>8</sup>

There is a parallel here between the way in which the content of an indexical expression is determined, and the characterization I gave of weak externalism. There is something that belongs to the expression, and that is intrinsic to the user of the expression – its character – that determines there to be a factor extrinsic to the user – a contextual parameter – that enters into the determination of the expression's content. It remains to apply this model to Mentalese in order to get a final picture of what weak externalism says about intentional content.

---

<sup>8</sup> OK, there's an indexical element in "child" – it's really hard to come up with examples that don't involve indexicals in some way or other. And it may be even harder than we thought, if subsequent discussion is correct.

Remember that, according to the CMM, Mentalese is a language in the fullest sense of the term. If so, then there is every reason to suppose that Mentalese contains both indexical and non-indexical expressions, just as natural languages do.<sup>9</sup> And in that case, we can distinguish expressions that have constant character from those that have variable character. The Mentalese term LOUISE ANTONY will have constant character; the Mentalese expression THAT PERSON KNITTING will, in virtue of the indexical component, have a variable character – it will pick out different knitters depending on the thinker’s location in space and time. Now we can encapsulate the disagreement between the weak internalist and the weak externalist about a term like WATER in this way. The weak internalist will say that WATER has a constant character – its satisfaction conditions are constant, and it delivers the same stuff in all contexts. The weak externalist will say that it has a variable character – its satisfaction conditions include a contextual parameter, so that the term will deliver different contents in different contexts. (Obviously, weak externalists need not agree about what the contextual parameter is – we’ll return to this point below.)

Weak internalists and weak externalists should disagree about twin-cases. A twin-case, remember, involves molecular duplicates who inhabit different regions of the same possible world. If there is no contextual factor built into the satisfaction conditions for a mental predicate like WATER, if WATER picks out the same stuff across contexts, then no two twins could have WATER terms that denote different stuff. The only way that the differences in their contexts – me on Earth, my twin on Twin-Earth – could make a difference to content is if, contrary to the weak internalist’s supposition, there is

---

<sup>9</sup> Joe Levine has even begun to think about what indexicals in the language of thought might be like. See his “Demonstrating in Mentalese” [\[give full ref\]](#)

some contextual parameter built into the character of WATER. On the weak externalist's view, twins of some sort are to be expected.

Now here's a question: is Fodor a weak internalist, or a weak externalist? Fodor himself explicitly allows the possibility of twins; I don't see, however, how he can, given his theory of content. If content is determined by the pattern of nomic dependencies between tokenings of mental symbols and instantiations of properties in the world, I simply don't see how molecular duplicates could differ in their intentional contents. Surely whatever laws subsume me subsume molecular duplicates of me. Whatever counterfactuals are true of me ought to be true of my twin as well. The only counterfactuals that will differentiate us, significantly, are those the expression of which make essential use of indexicals: counterfactuals like "if I were to be shown that this stuff had chemical properties A,B, and C, I would decline to call this stuff water." The same counterfactual, expressed in the same way and applied to my twin will yield the result that she and I will diverge in our dispositions to token WATER. But if we specify the same counterfactual conditions non-indexically, symmetry is restored. It will be true of both me and my twin that, shown that stuff S has chemical properties A, B, and C, we will decline to call stuff S WATER.

For reasons I've never completely understood, Fodor insists on his theory's ability to assign different extensions to twins' concepts. What he says is that there is a contextual parameter assumed in his account – that water is the stuff around here that's related to me in accordance with the asymmetric dependency condition. But this addition seems to me to be ad hoc, and not in the spirit of his account of intentional content. The parameter is not well-defined, and invites suspicions of indeterminacy. If

he embraced weak internalism, on the other hand, all he would have to give up is the conceptual possibility of twins. That would not be a particularly bitter pill to swallow, since he is already on record [[Elm and the Expert](#)] as holding that twins are nomologically impossible. But no matter. If he wants to build the contextual parameter in, then it turns out that he's simply a weak externalist. What I think of as the "pure" asymmetric dependency condition remains as an example of weak internalism.

What about Putnam? Obviously he is some sort of weak externalist. The question is, just what is the contextual parameter that he takes to enter into the character of concepts like WATER? In "The Meaning of 'Meaning,'" he claimed somewhat incautiously that natural kind terms contained an "indexical element." Tyler Burge objected that Putnam was wrong to say this, because "water" (or WATER) is not indexical in the way that "I" or "here" is indexical – it doesn't change its content depending on the context in which the thinker finds themselves. Putnam himself wants it to work out that if I travel to Twin-Earth, and say, upon arrival, "There's certainly a lot of water on this planet," that my term "water" retains its extension – it still refers to H<sub>2</sub>O. I used to agree with Burge's suggestion, but now it seems to be that Burge was not thinking of indexicals in a sufficiently abstract way. There's no reason, it now seems to me, that the relevant contextual parameter in the case of a term like "water" (or, for my purposes, a concept like WATER) could not be a historical one. The only trick is to figure out how history can get itself into the here and now of an individual's psychology.

Here is one way it might work: Imagine that Hermione is acquiring vocabulary items in her natural language, and suppose that today is the day that she acquires "arthritis." Perhaps she overhears her father telling someone that his uncle's arthritis

has gotten worse. Hermione may figure out that arthritis is some kind of disease, or she may not. She may reflect on the fact that her uncle is considerably older than she is, and form the belief that arthritis is something that older people have. She may think that arthritis is a kind of animal that her uncle keeps as a pet. It doesn't matter – all she needs to do in order to pick up the word arthritis – to add it to her lexicon of natural language expressions, so as to have it available for the expression of propositions – is to commit herself to meaning by “arthritis” whatever Dad meant by “arthritis.” Hermione thinks to herself, as it were, “I'm going to use that word ‘arthritis’ to refer to whatever Dad was referring to when he used it.” This much, I take it, follows from the Kripke/Putnam story about reference transmission.

Notice that this is not the same thing as Hermione's deciding to use “arthritis” as an abbreviation for the definite description, “whatever Dad means by ‘arthritis.’” I'm imagining that the demonstration involves what Kaplan calls the “dthat” operator. Hermione's thinking “Dwhatever Dad means by ‘arthritis’” puts her into contact with the thing that satisfies the descriptive condition, but it does not build the descriptive material itself into Hermione's thought. Demonstrations, on Kaplan's view, serve to introduce objects into propositions – they use descriptive material to get hold of the object, but once the object is grasped, the descriptive material falls away.

In this way, a current demonstration can put Hermione into contact with the past. Her connection to her Dad's use puts her into contact with the uses that her Dad keyed his own use to, and so forth, back to the original dubbing. The natural language word is, as it were, the tangible trace of the word's own history. Perhaps we could even think of the word itself as being an object extended in time. [\[What does Kaplan say? Check\]](#)

But now what happens inside Hermione's head? Hermione will want to think about arthritis, maybe learn something about it, like whether it's a disease, or a pet. When she does this, we want her to be thinking about the disease itself, and not merely about the natural language expression, "arthritis." For these purposes, it will be convenient for Hermione to introduce a new expression in Mentalese, ARTHRITIS, which she will use to refer to arthritis, and which can serve, henceforth, as the sense of the natural language expression "arthritis." This new Mentalese term does not refer to the natural language term "arthritis." Nor does it abbreviate the definite description, "the reference of the natural language word 'arthritis.'" Rather, the reference of ARTHRITIS is fixed by the demonstration to "arthritis." In this way, ARTHRITIS refers through that term to its reference.

It is part of the character of the new term ARTHRITIS – and of other similarly introduced terms – that it contains as a contextual parameter a demonstration of a natural language expression. Once the demonstration is made, the contextual parameter is filled in; from then on, there is no contextual variability in the reference of the Mentalese expression. But the fact that the contextual parameter is present at all allows us to explain how Putnamian twins could have thoughts with distinct contents. That's not to say that I necessarily agree with Putnam about the fundamental intuition. I've outlined one way a Mentalese expression might get its reference fixed, but there are others. Indeed, if the CMM is correct, then the semantics of at least a critical portion of Mentalese cannot depend, in the way I've described, on the representational properties of natural language. A fully interpreted Mentalese must be in place in order for conventional language to have emerged in the ancestral species, and for individual

human beings (or apes) to acquire it in their own personal lifetimes. But for that matter, Kripke and Putnam need this too. It's a crucial part of their stories about reference-fixation that speakers have intentions; intentional content cannot originate with the intentionality of language. So for a variety of reasons, the process I imagined for the reference fixation of Hermione's term ARTHRITIS cannot be the whole story.

I can think of at least two other possible ways it could go. First, natural language terms could always be given contents in just the way Kripke and Putnam say they are. (Surely proper names are introduced in this way.) All we would need in place is enough Mentalese repertoire to formulate the appropriate "dthat-" expressions. Since the descriptive material in such demonstrations falls away, there would be no question of defining the new expression in terms of Mentalese terms, nor of (this is actually equivalent) thinking of the properties used to specify the thing as essential to the thing. That's one way.

But secondly, there's always the possibility that some of the expressions of Mentalese have a weakly internalist semantics – that is, that they have constant characters. This is the way it would be if something like Fodor's asymmetric dependency account (understood in the way I've argued it ought to be, as lacking a contextual parameter) applies to some or all of the vocabulary of Mentalese. This would allow for concept nativism, if one wants it. (And I think there's lots of reasons why one should.<sup>10</sup>) What's specified innately, on this story, would be a psychological organization that effects a line-up between certain elements of a thinker's Mentalese terms with certain properties or kinds of things in the thinker's environment – we're born

---

<sup>10</sup> The literature is large. But see Fodor [\[various\]](#), Margolis [\[199?\]](#) and Antony [\[200?\]](#)

prepared to “track”<sup>11</sup> certain kinds of things, whether or not we actually run into them.

The relation between these Mentalese expressions and those of natural language is, of course, different than in the case of Mentalese expressions with the particular kind of variable character posited above. If I am born with a lexical item that tracks water – call it WATER – then the acquisition of my natural language expression, “water” must proceed differently. In this case, where I have a way of thinking about water antecedent to having acquired a natural language expression that means *water*, the acquisition will presumably involve the projection of a hypothesis, something like, “the stuff that ‘water’ refers to is WATER.” Importantly, this will be an empirical hypothesis, subject to error. If the natural language expression “water” refers to something different from what WATER refers to, then there may be a divergence between what I think and what I say. That seems to me to be a good result, and one that’s potentially useful for explaining things like the relation between “folk” and “scientific” concepts.

Notice that I’ve left it an open question what the reference of my (as we’re supposing) natively specified Mentalese expression WATER actually is. On this matter, I’m prepared to let the question turn on the nomological facts relevant to the asymmetric dependency condition (or whatever other weak internalist account might be the correct one for native expressions.) I’ve already committed myself to the view that Putnamian twins cannot vary with respect to the content of their WATER-thoughts, if WATER has constant character. It follows that, if there were such a thing as XYZ, then both my Mentalese term WATER and my twin’s term WATER have both H<sub>2</sub>O and XYZ

---

<sup>11</sup> In the sense of Loewer and Rey [\[199?\]](#)

in their extensions. That is, if WATER is a term with constant character for us, WATER must denote watery stuff. Now if it could happen, as Putnam's story goes, that each of us, upon learning a bit of chemistry, is disinclined to apply the word "water" to the watery stuff on the other's planet, this can be accommodated. The simplest thing to say is that each of us, upon acquiring our respective natural language terms, formed an incorrect hypothesis, viz., that something is the stuff called "water" just in case it is WATER. (Of course the reference to "water" has to be understood as indexed, in the way I've described, to different words for my twin and me. So the contents of our hypotheses were different.)

So this is my proposal: that a Mentalese expression can have either constant character, in which case its semantics is weakly internalist, or it has variable character, in which case its semantics is weakly externalist. If the latter, there is a contextual parameter that links the term either directly to something in the world (in the case of a dubbing) or else indirectly, via a natural language word. The elements we have in play are: content (always wide) and lexicon. We get the effect of sense by exploiting the lexical differences among vocabulary elements in the language of thought. And we can take advantage of the distinction between content and character to characterize two distinct ways in which the semantics for Mentalese items can be determined, with the possibility of reconciling weak internalism with weak externalism.

What does this mean for the "width" of psychology? I do not have space to review here all the arguments for the position that psychology should generalize over things in virtue of their internal constitution, but it is the position that I am prepared to defend. My suggestion is that psychology is interested in the character of thought, not

in its contents. Suffice to say that, unless strong externalism is true, there is an internal component to the determination of content. Given that there are guaranteed to be important generalizations that hold of thinkers in virtue of their sharing that internal component, there is going to be some version of psychology, with robust predictive and explanatory power, that will be narrow. Putnamian twins, “swamp-twins” and “vat-twins” (brains-in-vats molecularly identical to some embodied twins) will all share, in some sense, a psychology.

None of these pairs of twins, however, can be guaranteed to share contents. Whether twins share contents depends on two things: whether the thoughts in question have constant or variable character, and how cooperative their respective environments are. If, for example, I have a Mentalese expression that I’ve keyed to the reference of an ostended natural language term – getting the effect of a historical parameter within the character of the term – then my Swamp-twin will have the same Mentalese expression, but it will lack content. In general, if a Mentalese term requires the fixation of a contextual parameter, there will be the possibility of ungroundedness. My Swamp-twin is thinking – she just isn’t thinking anything.

This can provide some purchase on the problem of “empty concepts.” It can happen – and does, fairly often – that we fix our Mentalese expressions to natural language expressions that in fact lack reference. “Witch,” “phlogiston,” and, in my opinion “God” are all of this sort. Some philosophers<sup>12</sup> have argued that such cases provide an argument for the existence of really narrow content – for the position I called “solipsistic” internalism. But I see no reason to grant this. Content still seems to me to

---

<sup>12</sup> For example, Gabriel Segal [\[A Slim Book About Narrow Content\]](#)

be essentially relational; that means that the content of thought can be object-dependent. The existence of the thought depends on the tokening of the Mentalese item – whether it has content or not makes no difference to its role in psychological processing. My introspective certainty that there’s something I’m thinking can be accounted for by the fact that I am tokening a sentence in the language of thought – that such a thing is happening is something introspectively evident to me. So I can be confident that I am thinking – that I am the subject of a mental state – even if I cannot be confident that there is something I am thinking of. There is one kind of exception. Cartesian thoughts – reflexive thoughts in general – are guaranteed, by their character, to have some content. If a creature is thinking “I am here now” the mere tokening of the thought guarantees appropriate values for the contextual parameters that constitute the characters of the component expressions.

Finally, if psychology generalizes in terms of the character of Mentalese expressions, there is no need to posit a special kind of solipsistic content to account for the sense in which twins both “think the same thing” when each thinks something like “I’m ill.” Just as the notion of character gives us a way of capturing the sameness of meaning between two utterances of such a sentence by different speakers without having to posit a special kind of content both possess, the character of Mentalese expressions is sufficient to capture the similarity in thought between us.

Let me close by acknowledging one problem for the view. One of the arguments for positing a level of narrow content, between, on the one hand the lexicon and syntax of Mentalese, and wide content is that we want psychology to be able to generalize over individuals who are not twins – who are not molecular duplicates. The problem

with my suggestion is that we have no reason to expect that we will be able to identify Mentalese terms with the same character across individuals. We can be confident that twins will “spell” their Mentalese words the same way, and that their grammars will be the same. But individuals who are not similar down to the molecular level, individuals who are vastly more different than that, cannot be expected to realize Mentalese expressions in exactly the same way. What then, can be the interpersonal criterion of lexical sameness? In the case of expressions with variable character, we can get some purchase on the problem. These expressions can be recognized by their functional roles – they will be, in an important sense, part of the logical vocabulary of the thinker’s language of thought. But for the non-logical vocabulary, I can’t see what to say. It’s a problem that will have to wait for the next chapter.