The Content Program Through an Instrumentalist Lens*

Ori Simchen

Abstract

Theoretical representations in discussions surrounding the semantic significance of words and their analogs in thought should not be viewed under a realist interpretation as individually revealing what the represented items really are. Instead, they should be viewed under an instrumentalist interpretation as having other roles to play within their respective explanatory contexts. I consider some case studies for this broad methodological claim: theoretical representations of the semantic significance of words within semantics, theoretical representations of what determines the semantic significance of words within metasemantics – both under the auspices of the so-called new theory of reference – and theoretical representations of cognitive attitudes within the Representational Theory of Mind.

1 Introduction: Two Attitudes

Contemporary philosophy of language and mind is replete with appeals to content, both in philosophical semantics and elsewhere in the philosophy of mind broadly construed. Contents are theoretical representations of the significance of linguistic expressions and aspects of mental states and episodes. It is sometimes assumed that the notion of content is pre-theoretical. This is testimony to how entrenched content has become in philosophical discussions of linguistic and mental phenomena. But the idea that what accounts for the significance of linguistic expressions is also what is true or false, which is also what provides objects for cognitive attitudes such as beliefs, hopes, fears, and so on, is certainly not a pre-theoretical idea but a basic tenet of what might be called the *content program* in the philosophy of mind and language. The notion of content may have different roles to play in different explanatory settings. Whether or not it should ultimately be one and the same posit across

*To appear in *Synthese*.

distinct theoretical contexts is an interesting question I will not attempt to settle. My aim is to make an extended case for what I am going to call an *instrumentalist* interpretation of familiar swaths of theory within the content program.

Consider the following contrast between two sorts of attitude we adopt towards theoretical representations. The first is a *realist* attitude reserved for theoretical representations that are presumed to encapsulate or reveal what the represented items really are, however partially. This is how we often treat individual theoretical representations in the natural sciences, such as the representation of the sun as a sphere of hot plasma converting hydrogen into helium in astrophysics, or the representation of water as hydrogen hydroxide in physical chemistry, or the representation of gold as a transition metal. In such cases the individual representation is itself taken to reveal what the represented item really is at bottom, regardless of our more sophisticated opinions on the metaphysics of essence. Indeed, we naturally and intuitively adopt a realist attitude towards such theoretical representations *before* staking a position on whether essences are real or nominal.

In contrast to such cases, there are explanatory settings where a realist attitude toward theoretical representations seems clearly inappropriate, cases where the individual theoretical representation isn't entrusted with revealing the nature of the represented item but is rather called upon to perform some other theoretical task within the overall explanation. In the latter sort of case, we do not treat the representation under a realist attitude but treat it under an *instrumentalist* attitude instead.[1] Examples that come to mind here are the representation of real numbers as Dedekind cuts in the foundations of mathematics, or the representation of gold as a monetary standard in economics, or the representation of the meaning of a sentence as a set of indices in formal semantics. In such cases we do not presume that the representation should itself disclose what the represented item really is. We don't expect, for example, that a certain set of indices should capture what the meaning of 'There is a bag of potatoes in my pantry' really is.[2] Whatever that meaning is, it surely isn't *that*. Rather, the set of indices represents the meaning within a broader theoretical context of formal semantic explanation that includes generating truth-conditions for whole sentences in a way that comports with structural requirements

[1]For a discussion of this attitudinal contrast, see my (2019). I do not delve into individuation conditions for theoretical representations here, leaving the matter deliberately open-ended. I assume that we represent things within our theories and that what ensues are theoretical representations of those things. I also do not delve into individuation conditions for theories. For present purposes bits of theory doing explanatory work within a broader theoretical setting count as theoretical representations.

[2]The point is familiar to semanticists and illustrated nicely in the opening passages of the classic Heim and Kratzer (1998).

introduced by generative syntax. Likewise, and in a very different setting, we don't suppose that the Dedekind cut tells us what the real number really is; we assume, rather, that the real is represented by – or "constructed as" – a Dedekind cut of rationals.

Attitudinal instrumentalism and realism should be distinguished from traditional instrumentalism and realism. Traditional instrumentalism in the philosophy of science is a position that concerns in the first instance claims about unobservables. Very roughly, the traditional instrumentalist proclaims that terms for unobservables do not stand for anything because there are no such things to stand for – terms for unobservables are not representational. Applied to specific claims relevant to the topic of this paper, a traditional instrumentalist about claims concerning mental phenomena, for example, holds that the terms that purport to refer to such phenomena do not because, strictly speaking, there are no such things – terms for mental states and episodes do not really represent. In contra-distinction to a traditional instrumentalist stance, instrumentalism in the sense relevant to this paper is an attitude towards theoretical representations according to which said representations do not themselves reveal – individually, as it were – what the represented things really are.[3] An instrumentalist attitude in the relevant sense presupposes that the relevant theoretical representations do in fact represent. Being a standard for pre-20$^{\text{th}}$ century monetary systems can represent gold theoretically even if it doesn't disclose the nature of the represented substance. Or so claims the instrumentalist in the relevant sense, and common sense clearly agrees here.

The distinction between a realist and an instrumentalist attitude towards theoretical representations has far-reaching implications for philosophical theorizing. Philosophy is a broadly theoretical enterprise where it is often assumed that bits of theory should be taken in a realistic spirit by default.[4] Examples of this way of thinking are too common to survey here, but let us call a default adoption of a realist attitude towards theoretical representations in philosophy *metaphilosophical realism*. Its opposite, *metaphilosophical instrumentalism*, is the more liberal approach according to which an instrumentalist attitude towards theoretical representations in philosophy may be (and often is) warranted. Metaphilosophical instrumentalism gives us freedom to maneuver among various first-order theoretical options in philosophy with relative ease. Without a default commitment to treating philosophical theoretical representations as revealing the nature of what they represent individually, we

[3]To speak of theoretical representations "themselves" or "individually" revealing what the represented things really are is surely an idealization but a useful one for the purpose of drawing the very real attitudinal contrast under consideration.

[4]See my (2019) for an extended articulation of this claim.

can pause and ask what explanatory work those representations are called upon to do. My aim here is to explore central aspects of the content program through an instrumentalist lens as part of a more general effort to promote metaphilosophical instrumentalism in the philosophy of language and mind.

A methodological point is in order before we proceed. I am not going to argue directly against the adoption of a realist attitude and in favor of the adoption of an instrumentalist attitude towards theoretical representations within the content program. My aim, rather, is to illustrate the explanatory advantage of adopting an instrumentalist stance by focusing on central aspects of the program and viewing them from an outlook that is superior to the common realist one. In other work I've proposed guidelines for warrant in adopting a realist attitude towards theoretical representations.[5] The basic idea is to take for granted such warrant for uncontroversial cases outside philosophy – geometrical solids representing minerals in crystallography, for example – identify central aspects of those cases, and apply them as necessary conditions for such warrant in the more controversial cases of philosophical theorizing. Implementing such a strategy here would require showing that some of the hypothesized necessary conditions for warrant in adopting a realist attitude towards the target representations fail. For example, it appears that in the uncontroversial cases of a warranted realist attitude the fact of representation itself falls within the purview of the surrounding theory. That water is theoretically represented as being hydrogen hydroxide, for example, is covered by a broad physical-chemical account that ties together macro and micro features of the substance. Nothing like this can be said about many familiar theoretical representations in philosophy, such as a possible world representation of a distinct possibility within the metaphysics of modality. The surrounding theory here leaves us none the wiser as to *why* the possible world, however ultimately construed, should theoretically represent what is really possible in a given case. In any case, my aim in this paper is more modest. I propose to give an instrumentalist interpretation of basic aspects of the content program by focusing on familiar foundational work within it. I aim to show that an instrumentalist stance enjoys greater plausibility than its realist rival when it comes to theoretical representations of meaning, of meaning-determination, and of cognitive attitudes within the content program. For reasons that fall outside the scope of this paper, the swaths of theory to be discussed are commonly viewed under a realist interpretation. This is a fundamental oversight about the explanatory work and achievement of familiar aspects of the content program, as we will see.

[5]See my (2019: Sec. 5).

2 Attitudinal Instrumentalism and Semantic Significance

The discussions surrounding content within contemporary philosophy of language and mind are generally framed by engagement, both positive and negative, with the so-called new theory of reference of the 1970s. The influence of the new theorists' work on subsequent philosophy has been nothing short of profound. Before this work it was commonly assumed that the terms of language and thought are associated with conditions that fix what the terms are about. For example, singular terms such as 'Elizabeth Warren' were thought to be associated with conditions specifying the term's referent; general terms such as 'Democrat' or 'progressive' were thought to be associated with conditions specifying the term's range of application. Treating referents for singular terms and ranges of applications for general terms as the terms' extensions, endowment with semantic content was thought to consist in the association of terms with extension-fixing conditions. And knowledge of such conditions was thought to comprise semantic competence.

All this changed following the work of Donnellan (1966, 1970), Putnam (1975), Kripke (1980), and others, work that convinced philosophers that the old view of semantic endowment and competence was wrong not only on this or that detail but fundamentally. An average speaker may be proficient with 'Cicero' and use it to refer to Cicero without being able to identify Cicero beyond being some famous Roman orator, or being some famous Roman, or perhaps even just being some famous guy. An average speaker may be proficient with 'elm' and 'beech' and use them to specify the elms and the beeches, respectively, without having in her cognitive possession a condition associated with each term that allows her to distinguish the elms from the beeches. Pre-1750 Oscar here on Earth may think a thought he would express by saying 'Water is abundant' while his pre-1750 *Doppelgänger* on Twin-Earth may think a thought *he* would express by saying 'Water is abundant'. Oscar and his twin are in every relevant way the same. But the stuff surrounding Oscar is H_2O while the stuff surrounding his twin is some alien but superficially similar stuff XYZ. So Oscar's thought and his term 'water' are about H_2O; his twin's thought and *his* term 'water' are about XYZ. Such examples may be multiplied as needed. The negative upshot is that semantic endowment isn't the association of terms with extension-fixing conditions and semantic competence isn't knowledge of such conditions. But in the wake of this important body of work the question remains how to think positively about semantic significance.[6]

Putnam's classic "The Meaning of 'Meaning'" (1975), a cornerstone of the new

[6]I set aside the question of narrow content. See my (2004) and (forthcoming) for a discussion of narrow content in relation to the explanatory achievements of the new theory of reference.

theory of reference, exhibits heightened sensitivity to methodological issues that lie at the heart of the theoretical study of meaning and are easily overlooked. There is a clear instrumentalist streak (in our sense) running through Putnam's thinking about semantic significance:

> Briefly, my proposal is to define "meaning" not by picking out an object which will be identified with the meaning (although that might be done in the usual set-theoretic style if one insists), but by specifying a normal form (or, rather, a *type* of normal form) for the description of meaning. If we know what a "normal form description" of the meaning of a word should be, then, as far as I am concerned, we know what meaning *is* in any scientifically interesting sense. (190)

And a little later, regarding a normal form description of the meaning of 'water' that includes specifying H_2O as the term's extension, alongside other parameters such as syntactic marker, semantic marker, and stereotype, we read:

> [T]his does *not* mean that knowledge of the fact that water is H_2O is being imputed to the individual speaker or even to the society. It means that (*we* say) the extension of the term 'water' as *they* (the speakers in question) use it is *in fact* H_2O. (191)

The picture that emerges from such passages is that in characterizing the meaning of 'water' we are not offering a theoretical identification that would tell us what that meaning is in the way imagined perhaps by the metaphysician of meaning and modeled after theoretical identifications in natural science, such as water being H_2O. We can say what we need to say theoretically about meaning by giving a normal form description of the meaning of the term. We do this without providing a theoretical representation that would itself be entrusted to reveal what that meaning really is, as one might capture theoretically the constitution of a substance in physical chemistry by representing it as a certain chemical compound. Putnam's talk of not picking out some object to be the meaning while leaving room for its set-theoretical representation ("although that might be done in the usual set-theoretic style if one insists") is reminiscent of a similar point urged by Lewis regarding a different but related explanatory enterprise. The point merits a brief digression on Lewis's own instrumentalist predilections in the study of language and thought.

Lewis (1975) identifies language use in a given population as the existence of a convention of truthfulness and trust in £, where £ is a mapping of strings of signs onto sets of possible worlds (indices). To be truthful in £ is to try not to issue a false sentence – construed as a sentence σ for which $@ \notin £(\sigma)$, where @ is the actual

world – and to be trusting in £ is to impute truthfulness to others. This convention of truthfulness and trust in £ is a kind of regularity in linguistic behavior within the population, a regularity of not uttering sentences believed to be false and expecting others to do the same. Lewis's six conditions for such a regularity to amount to a convention need not concern us. What is of present concern, however, are the levels upon levels of theoretical representation within this work, none of which needs to be treated under a realist attitude. First and foremost, we have the language used by the population represented as a function £ from strings of signs to sets of possible worlds. This function is represented set-theoretically as a set of ordered pairs. The strings of signs making up the first members of those pairs represent sentences; the sets of possible worlds making up the second members of those pairs represent sentence-meanings ("propositions"). A sentence represented by the string σ being true or false is represented by the actuality-representing @ being a member or not of the set $£(\sigma)$.

Lewis considers many objections to his theory of language use and offers detailed replies. An important objection runs as follows:

> Unless a language user is also a set-theorist, he cannot expect his fellows to conform to a regularity of truthfulness and trust in a certain language £. For to conform to this regularity is to bear a relation to a certain esoteric entity: a set of ordered pairs of sequences of sound-types or of mark-types and sets of possible worlds.... The common man has no concept of any such entity. Hence he can have no expectations regarding such an entity. (Lewis 1975: 24-5)

Insofar as a language is a pairing of sentences with their meanings, it would appear that a language user would need to have some cognitive rapport with the pairing in question, the set-theoretical entity. This would require, in turn, rapport with sentences construed as strings, which are further set-theoretical entities, and rapport with the meanings of sentences, construed as yet further set-theoretical entities. Does it not follow that the language user would need to be cognitively *en rapport* with an elaborate set-theoretical construction? Lewis replies:

> The common man need not have any concept of £ in order to expect his fellows to be truthful and trusting in £. He need only have suitable particular expectations about how they might act, and how they might form beliefs, in various situations.... He expects them to conform to a regularity of truthfulness and trust in £ if any particular activity or belief-formation that would fit his expectations would fall under what *we* – but not *he* – could describe as conformity to that regularity. (25)

£ is a formal object representing a language, just as the non-membership of @ in £(σ) represents falsity for the sentence represented by the string σ. It isn't a representation that we should be treating as itself telling us what a language is in the most demanding sense, or what issuing a meaningful sentence is, or what interpretation amounts to. Nor does the representation of sentences as strings of signs in the set-theoretical sense tell us what sentences themselves really are at bottom. Whatever sentences are, it's unlikely they're sets. Nor does the representation of the meaning of a given sentence as a set of possible worlds tell us what that meaning really is. Lewis's theory of language use deploys various theoretical representations in an effort to shed explanatory light on a wide-ranging phenomenon. Treating those representations individually in a realistic spirit is no part of this explanatory endeavor.

Similarly, those who are expecting the new theory of reference to represent meanings in a way that reveals what those meanings really are at bottom are bound to come away disappointed. One of the easily missed lessons of the new theory of reference is that we need not be beholden to a conception whereby representations of meaning are themselves revelatory of the nature of the pre-theoretical subject matter. A question often raised in reaction to the now familiar arguments in Donnellan (1966, 1970), Kripke (1980), and Putnam (1975) is: "So what *is* the meaning of a referentially used description, or of a proper name, or of a kind term?" Such questions belie a misunderstanding. The correct response to such questions is: "For what explanatory purpose, exactly?" There may be good theoretical reasons to treat descriptions and proper names, for example, as denoting individuals for formal semantic purposes. Those denotations are no more meant to tell us what the meanings of those expressions really are than the representation of the meaning of a sentence as a set of indices is meant to tell what the meaning of the sentence really is. Specific theoretical representations of meaning – that a particular description denotes (in the formal semantic sense) a particular individual, for example – are not themselves entrusted to reveal the nature of the pre-theoretical subject matter – what the meaning of the description really is.

3 Attitudinal Instrumentalism and Metasemantics

The new theory of reference shouldn't be taken as offering theoretical identifications, along the lines of gold being the element with atomic number 79 or water being H_2O, when it comes to semantic significance. Treating the theory's pronouncements as purporting to reveal what semantic significance really is in given cases misconstrues the theory's scope and achievement. But philosophers have also looked to this body of work for how to think of the determination of semantic facts, specifically

for how language and thought "hook onto the world". Sub-sentential expressions make various contributions to the truth-conditions of the sentences in which they partake. Semantics specifies those contributions whereas metasemantics inquires after how those particular contributions get determined. Before the advent of the new theory of reference the distinction between semantics and metasemantics was not clearly drawn. Once semantic contributions of sub-sentential expressions to truth-conditions of sentences became more austere – individuals in the case of proper names, for example – a separate line of inquiry could emerge targeting the determinants of semantic relations – how, for example, a particular name came to name a particular individual. Such metasemantic matters, on a common way of looking at them, form a branch of metaphysical inquiry, and the question I turn to next is how to regard theoretical representations within this branch of inquiry in the wake of the new theory of reference. I aim to show that an instrumentalist attitude can play an important role within a defensible variant of the metasemantic picture handed down to us by the new theorists. The situation here is complicated by certain entrenched ways the theory has often been received, but I will illustrate the need for an instrumentalist attitude in appreciating the work's real explanatory power.

The need for an instrumentalist attitude towards theoretical representations in metasemantics can be usefully illustrated by considering some of the details surrounding extension-fixing for general terms. An important component of Putnam's (1975) overall conception, for example, is the suggestion that extension-fixing for a general term such as 'water' is crucially demonstrative (or "indexical"). After presenting his sociolinguistic hypothesis of the division of linguistic labor, Putnam (1975: 148-149) compares two alternative characterizations of demonstrative extension-fixing for 'water' and opts for the second:

(1′) for any world w and any x in w, x is water just in case x is the same_{L} as the referent of 'this' in w,

(2′) for any world w and any x in w, x is water just in case x is the same_{L} as the referent of 'this' in actuality.

According to (1′), to be water in any world is to be the same_{L} as a sample of local watery stuff – clear, potable liquid, raining from the sky, filling the lakes, etc. – demonstratively referred to in the world in question. According to (2′), to be water in any world is to be the same_{L} as a sample of watery stuff demonstratively referred to in the actual world. $\text{Sameness}_{\text{L}}$ is a cross-world relation of sameness for liquids that prominently features chemical composition. It is (2′) rather than (1′) that comports with our referential intentions as users of 'water', says Putnam. We intend 'water'

to apply to all and only samples of *water*, i.e. H_2O plus or minus impurities, at *any* world. A world where H_2O is entirely absent and a superficially similar alien substance XYZ occupies the role played by H_2O in actuality is a world where there is no water at all despite the abundance of the superficial look-alike. That 'water' applies to water and only to water in any possible world is characterized by Putnam, using Kripke's terminology, as *rigidity*[7]:

> If we extend the notion of rigidity to substance names, then we may express Kripke's theory and mine by saying that the term 'water' is *rigid*. The rigidity of the term 'water' follows from the fact that when I give the ostensive definition "this (liquid) is water" I intend (2′) and not (1′). (149)

Speakers wield 'water' with the intention to specify anything, in any possible world, relevantly similar to *this* (as deployed in the presence of a paradigm actual sample of water). The rigidity effect appears to depend on the behavior of the implicated sameness relation.[8]

[7]Kripke (1980: 48) characterizes rigidity for singular terms as the designation of the same individual at every world in which the individual exists. How to extend the notion of rigidity to general terms such as 'water' has spawned a sizable secondary literature. Let us assume, however, that for a general term to be rigid is for the term to apply to an individual member or sample of an associated kind at any world in which the individual or sample exists. See Gòmez-Torrente (2006) for a discussion of this construal of rigidity for general terms.

[8]To see this, consider extension-fixing for a non-rigid general term along Putnamian lines. What my favorite color happens to be might have been different, so the complex singular term 'my favorite color' is presumably non-rigid – it doesn't designate the same color at every world in which the color exists. Accordingly, the general term 'sample of my favorite color' should presumably come out non-rigid as well in the sense given in the previous footnote. Let $\text{sameness}_{\text{FC}}$ be the relation among samples of color such that o is the same_{FC} as o' just in case each is a sample of my favorite color in its respective world. (Such a cross-world relation prominently features, let us suppose, psychological goings-on associated with certain types of response-dependence.) If o and o' are worldmates, then they sample the single color that happens to be my favorite in their shared world, whereas if they're not worldmates, each samples my favorite color in its respective world. We can characterize extension-fixing for 'sample of my favorite color' as follows:

(3′) for any world w and any x in w, x is a sample of my favorite color just in case x is the same_{FC} as the referent of 'this' in actuality.

The structure of (3′) is the same as that of (2′). And yet 'water' is rigid whereas 'sample of my favorite color' is not. A sample of blue in the actual world is a sample of my favorite color, blue, whereas that very sample in a world in which my favorite color is red will not be a sample of my favorite color in that world. So 'sample of my favorite color' will not apply to that sample at every world in which the sample exists.

Be that as it may, the account we've been considering is part of an overall metasemantic story offered to counter a traditional view according to which the terms we use in language and thought are associated with extension-fixing conditions knowledge of which accounts for semantic competence. An important consequence of the older view is that it makes fixity of subject matter across radical changes in theory and belief difficult to attain. As such, the older view isn't as well equipped to handle theoretical progress. On the older view, in order for a contemporary theoretically-informed speaker and an ancient speaker using a substance-term to be on the same page when it comes to the substance at issue, the conditions associated with the term in each of their mouths are required to specify one and the same substance. This is unlikely, to say the least. The belief-set of contemporary speakers when it comes to water is quite different from the corresponding belief-set of ancient speakers. But on the account of extension-fixing we've been discussing, speakers intend to use 'water' for anything relevantly similar to paradigmatic samples. Insofar as contemporary and ancient samples are samples of the same substance, the term 'water' in the mouth of contemporary speakers can have the same range of application as the term in the mouth of ancient speakers despite significant disparities in beliefs and theory. What counts as relevant similarity for water is something that inquiry homes in on in the fullness of time. The important point is that according to the new view what counts as water doesn't pose an undue burden on speakers and thinkers about the substance.

We may construe Putnam's endorsement of $(2')$ over $(1')$ as filling in some of the implications of speakers' referential intentions in light of this last point. How to think of referential intentions is vexed.[9] Referential intention attribution is intention attribution, which is a type of cognitive attitude attribution. We shouldn't expect the representations deployed in characterizing cognitive attitudes to reveal the nature of the represented attitudes at this early stage of inquiry into such cognitive matters.[10] And we certainly shouldn't expect such representations to reveal the nature of extra-cognitive matters of fact such as the individuation conditions for substances. In characterizing referential intentions, 'same_{L}' represents whatever relevant similarity for liquids happens to be by the lights of the attribu*tor*, the theorist, not the attribu*tee*. There is considerable metaphysical controversy over how to think of relevant similarity for substances, but the outcome of this controversy is largely irrelevant to the proposed metasemantics for 'water'. The metasemantics stands on its own even in the absence of a satisfying metaphysical story about sameness for

[9]See my (2012: Ch. 3) for an extended discussion of the issue.

[10]See next section for further discussion of this point.

substances.[11] Sameness$_{\text{L}}$ is invoked here as part of a theoretical capture of speakers' referential intentions. We are told that the average speaker deploys 'water' with the intention to specify anything relevantly similar – i.e. anything that is the same as befitting liquids – to samples in the speaker's environment. Think of sameness$_{\text{L}}$ in the characterization of an average speaker's referential intention along the lines of £ as it figures in Lewis's characterization of language use discussed in the previous section. Lewis characterizes language use in terms of an attitude of expecting fellow speakers to be truthful in £ *inter alia*, where £ is an elaborate set-theoretical construction. The relevant objection considered by Lewis is that the average speaker cannot be expected to be able to form attitudes towards £ – such attitudes would require the speaker to be a set-theorist. Lewis's reply to the objection is that it isn't incumbent on the agent of the attitude to be capable of characterizing the attitude by appealing to £. Such an appeal is something we theorists do. To suppose that £ should be taken as a realistic theoretical representation of an aspect of the cognitive attitude of trust in a language is to misunderstand its explanatory role in the overall theory.[12] Along similar instrumentalist lines, Putnam's proposal that speakers intend their 'water' to pick out anything that is the same$_{\text{L}}$ as actual samples of water can be apt without making unreasonable predictions under a realist interpretation

[11]For a contrasting view, see Häggqvist and Wikforss (2018), who hold that metaphysical troubles concerning what counts as a given substance have dire consequences for the metasemantic picture outlined here. See also Needham (2017), who takes the notion of demonstrative extension-fixing for substance terms to be of limited reach. These discussions seem to miss relevant aspects of Putnam's metasemantic proposal under an instrumentalist interpretation. In characterizing the relevant referential intention, the role of the demonstrative pronoun 'this' is simply to highlight the environmental aspect of extension-fixing for a substance term by whichever means – that the term is intended for stuff "around here". Whether or not demonstrative reference plays any role in how substances are in fact named is beside the metasemantic point; the choice of particular semantic means is incidental. This choice should also not be taken to imply that mastery with 'water' depends on antecedent mastery with 'this', an empirical matter to be settled separately.

[12]For yet another striking example of Lewisian attitudinal instrumentalism in the philosophy of mind and language, consider Lewis's (1979) view according to which believing is represented as the self-ascription of a certain property, which is construed as self-location *within* the property, which is construed, in turn, as a set of possibilia. Such theoretical representations of the attitudes are clearly not to be taken under a realist interpretation. Lewis's account of the attitudes as the self-ascription of properties is not to be mistaken for the misbegotten idea that the cognizing agent is somehow invariably the topic of all her attitudes. On this point, see also an interesting discussion by Nolan (2006). At first blush, Nolan's critique appears to target the inevitable self-involvingness of attitudes under the Lewisian scheme, which would presuppose an independently unmotivated realist interpretation of the deliveries of the theory. But on a closer look, Nolan's complaint is the more subtle point that Lewis's failure to draw a distinction between self-involving and selfless attitudes is a shortcoming of the framework in its express explanatory aims, which does seem like a fair point.

of what goes on within speakers' mentality as they deploy the term. Moreover, just as controversy surrounding the metaphysics of sets is beside the theoretical point when it comes to Lewis's appeal to £ within his theory of language use, controversy surrounding what counts as the same substance is beside the theoretical point when it comes to Putnam's appeal to $\text{sameness}_{\text{L}}$ within his proposed metasemantics. Differences among various metaphysical takes on sets – platonism vs. nominalism, for example – are largely irrelevant when it comes to the explanatory achievement of Lewis's theory even if it utilizes sets. Differences among various metaphysical takes on sameness for substances – realism vs. nominalism about natural kinds, for example – are also irrelevant when it comes to utilizing sameness for substances within the target metasemantics.

4 Attitudinal Instrumentalism and RTM

Let us turn to the deployment of content within the metaphysics of cognitive attitudes. We saw that positing contents need not be committed to the idea that such representations of semantic significance are themselves revelatory of the nature of what they represent. We also saw that positing a particular demonstrative-cum-comparative structure for referential intentions need not be taken to reveal the underlying nature of the mental set required for acquisition of and proficiency with general terms. We will now see that positing certain mental particulars with a syntax and semantics as relata for cognitive states and episodes need not be taken to reveal what it is to be in such states and undergo such episodes. And yet the entire system of such theoretical representations may reveal something important about the nature of cognition. To fix on an image here, think of Mendel's theory of inheritance and his notion of "factor" as the unit of inheritance in formulating his principles. The Mendelian factor represents the gene, with its different "forms" representing different allelles. But we wouldn't say that the Mendelian factor individually reveals what the represented gene really is. In the course of subsequent genetic inquiry that story gets filled in, but Mendelian factors represented genes within a larger theoretical context that doesn't include delving deeper into what the represented genes really are. Such explanatory holism (for lack of a better term) is especially important to keep in mind when considering leading theories in the metaphysics of cognitive attitudes, views according to which being in a cognitive state such as believing that p consists in bearing a certain relation to something – a "mental representation" – that carries the semantic content that p.

The most developed and influential of those views is Fodor's Representational Theory of Mind (RTM). Fodor summarizes RTM as follows:

> At the heart of the theory is the postulation of a language of thought: an infinite set of 'mental representations' which function both as the immediate objects of propositional attitudes and as the domains of mental processes. More precisely, RTM is the conjunction of two claims:
>
> > *Claim 1* (the nature of propositional attitudes):
> > For any organism O, and any attitude A towards the proposition that P, there is a ("computational"/"functional") relation R and a mental representation MP such that
> > MP means that P, and
> > O has A iff O bears R to MP.
> >
> > ...
> >
> > *Claim 2* (the nature of mental processes):
> > Mental processes are causal sequences of tokenings of mental representations. (1987: 17)

RTM has been subjected to a host of criticisms and subsequent refinements over the years. The general tenor of Fodor's overall position is an abductive "only game in town" stance. What comparable alternative might there be, given what we aim to explain? Indeed, RTM has unmistakable reach when compared with some of its alternatives. How can we understand, for example, what it is for someone to believe the universe began with the Big Bang without positing a relation between the believer and something that behaves like the sentence 'The universe began with the Big Bang', with an attendant syntax and semantics? Such a posit easily figures in further explanations of systematicity in the agent's thinking. It can help explain why, for example, believing the universe began with the Big Bang plausibly gives rise to the belief that the universe had a beginning, that the Big Bang was in the past, and much else besides. Alternative accounts that seek to avoid such posits are in a difficult spot.

One such account is Marcus's (1990) dispositionalism. Marcus contrasts "language-centered" theories of belief such as RTM with the "object-centered" accounts she favors. On her view, an agent believes that p just in case under various agent-centered circumstances the agent is disposed to act as if the state of affairs that p obtains. We might understand "disposition to act" on this account as a way station in the quest after something more comprehensive and theoretically satisfying down the road of inquiry. Dispositionalism might seem like a reasonable reaction to certain perceived shortcomings of language-centered accounts, such as the construal of beliefs for nonlinguistic or prelinguistic creatures as relations to sentence-like things, and the construal of Kripke's (1979) logically blameless Peter believing Paderewski

had musical talent while believing Paderewski had no musical talent as someone who believes a contradiction. Nevertheless, a dispositional story will be hard pressed to distinguish believing the universe began with the Big Bang from believing the universe will end with the Big Crunch, or anything else without obvious ramifications for behavior. This is so unless "disposition to act" on the right hand side of the proposed account includes linguistic behavior. But then it seems reasonable to want to know *why* the disposition to utter 'The universe began with the Big Bang' is correlated with believing the universe began with the Big Bang, while the disposition to utter 'The universe will end with the Big Crunch' is correlated with believing the universe will end with the Big Crunch. It is a tall order. RTM, on the other hand, can enlist the gamut of syntax and semantics to distinguish such beliefs. The dispositionalist story, by comparison, seems ill equipped to track such cognitive fineness of grain.

A shortcoming of theories such as RTM not discussed by Marcus directly is the general matter of sheer theoretical overreaching, given the paucity of the evidence. Such theories seem to overreach empirically, which can easily incline critics to dismiss them as fanciful. Language-like intermediaries for cognitive attitudes can easily seem to emanate from theoretical wishful thinking. One can accept such criticism and yet wonder whether there isn't a less radical alternative to doing away with language-like intermediaries altogether in the metaphysics of cognitive attitudes.[13]

Fodor (1975) considers RTM's representations of cognitive states and episodes in terms of relations to such intermediaries as revealing the very nature of those states and episodes:

> To have a certain propositional attitude is to be in a certain relation to an internal representation. That is, for each of the (typically infinitely many) propositional attitudes that an organism can entertain, there exist an internal representation and a relation such that being in that relation

[13]In previous work I try to counteract the "only game in town" idea by developing a template for an alternative I call Cognitive Relations Theory (CRT). The theory does away with linguistically structured mental representations altogether. It views cognitive attitudes such as hunting, wanting, and worshipping, but also believing and the rest of the so-called propositional attitudes, as in the first instance direct relations to objects such as lions, sloops, and people. It is thus very clearly an object-centered theory in Marcus's sense. CRT treats specific attitudes – cognitive states directed at particular things such as Ernst hunting a particular lion or Ralph believing Ortcutt in particular to be a spy – as primary in the order of explanation, and conceives of such attitudes as putting agents in direct, unmediated contact with their objects without language-like intermediaries. For non-specific ("generic") attitudes such as hunting a lion but no lion in particular, or believing every spy is dangerous without believing any spy in particular to be such, CRT goes subjunctive. For Ralph to believe every spy is dangerous, for example, is roughly for the following to obtain: Had Ralph believed anything in particular to be a spy given his actual mental set, Ralph would believe that thing to be dangerous. For further discussion, see my (2012: Ch. 5).

> to that representation is nomologically necessary and sufficient for (or nomologically identical to) having the propositional attitude. The least that an empirically adequate cognitive psychology is therefore required to do is to specify, for each propositional attitude, the internal representation and the relation which, in this sense, correspond to it. Attitudes to propositions are, to that extent, 'reduced' to attitudes to formulae, though the formulae are couched in a proprietary inner code. (198)

When Fodor describes himself as an intentional realist and takes RTM to be the only game in town, the upshot for him is that the representation of a given cognitive fact within the theory reveals the very nature of this fact. But this need not be the only option. Might not RTM as a whole reveal something important about the nature of cognition without each of its theoretical representations doing so individually?

In an illuminating discussion of sententialism about belief reports – the view that the truth-condition for 'A believes that p' is the agent bearing a certain relation of "believing-true" to the complement clause – Quine (1956) writes:

> This semantical reformulation is not, of course, intended to suggest that the subject of the propositional attitude speaks the language of the quotation, or any language. We may treat a mouse's fear of a cat as his fearing true a certain English sentence. This is unnatural without being therefore wrong. It is a little like describing a prehistoric ocean current as clockwise. (186)

To say that the mouse fears the cat is about to pounce is to say, according to the sententialist, that the mouse fears-true the English sentence 'The cat is about to pounce'. This, we are told, no more requires us to attribute the speaking of English to the mouse than describing a prehistoric ocean current as clockwise requires us to postulate some interesting relation between the prehistoric ocean and clocks. The theorist is offering truth-conditions for the likes of 'The mouse fears the cat is about to pounce'. It is no part of the theoretical effort here to get deeper into the cognitive facts being reported beyond delivering the right truth-conditions for the reports. However the mouse is tracking the cat's readiness to pounce, its fact-directed fear is the truth-condition for 'The mouse fears the cat is about to pounce'. And this fact-directed fear may be construed by the theorist as a relation to an English sentence.

Switching from sententialism in the semantics of attitude reports to the metaphysics of the attitudes themselves, we can deny the realist interpretation of the deliveries of a theory such as RTM with respect to our mouse's fear along analogous lines. RTM would represent the fear in question as a functional relation R obtaining between the mouse and a mental representation MP that means that the cat is about

to pounce. But we need not think of this theoretical representation as itself revealing the nature of the represented fear. What we aim to lay bare with RTM is a general cognitive architecture for the mouse. We might consider the mouse's fear that the cat is about to pounce as interestingly related to the more generic fear that something or other is about to pounce. We might characterize the mouse's fear as interestingly related to the mouse's belief that there is a cat in front of it, or to the belief that there is a cat in front of it about to pounce, or perhaps to the highly specific belief that *this very cat* (the proximal source of pheromones) is about to pounce. What the individual attitudes really are at bottom, the nodes in the overall structure of mouse mentality, need not be part of this explanatory enterprise. In a similar vein, we may represent a substance with a plastic model of its molecular structure, consisting of spheres and connecting rods that in turn represent atoms and chemical bonds. The model as a whole may reveal the molecular nature of the represented substance without the spheres and rods individually revealing the natures of the represented atoms and bonds.

These are early days of cognitive theorizing. Perhaps instead of starting to theorize from scratch under various idealizations in order to avoid empirical overreaching, we ought to take a more cautious attitude towards what we've already got.[14] Such conservatism in theory-choice, coupled with an instrumentalist stance in theory-interpretation, need not be viewed as abandoning the main tenet of Fodorian intentional realism. There *is* a fact of the matter about cognitive states and processes, but it is revealed by larger swaths of theory rather than by individual representations of particular manifestations of mentality. A healthy instrumentalist attitude in the metaphysics of mind allows us not to throw out the baby with the bath water here.[15]

Fodor (1975) famously introduces his Language of Thought Hypothesis with a critique of reductionism. He discusses the special sciences vis-á-vis physics; economics, in particular, is offered as a salient example of an implausible reduction to physics. But there is a general lesson regarding special scientific explanation that

[14]See previous footnote. CRT proceeds by way of the idealization that complement clauses for attitude reports are fully regimentable without loss into first-order logic. The theory is then put to a formal test via a proof, within a quantificational extension of Lewis's counterfactual logic VC, that the set comprising the deliveries of CRT with respect to an omniscient believer – a believer whose belief states are reported with complement clauses that themselves comprise a consistent set of sentences – is consistent. See my (2012: Appendix II).

[15]A promising reinterpretation of RTM along such lines is provided by Rescorla (2020) whereby mental representations are conceived as abstracta that theoretically represent actual representational capacities. Those abstracta are clearly not meant to disclose what the represented representational capacities really are at bottom. A system of such abstracta, however, can constitute a genuine step forward within the metaphysics of mind.

Fodor fails to heed. Economists can insist that there is a fact of the matter about the nature of monetary systems vis-á-vis the gold standard, or about currencies in formulating Gresham's Law, without having anything interesting to say about the underlying nature of gold or the underlying nature of ratios of commodity value to nominal value for currencies. Something similar, I claim, can be said about representing a cognitive state such as a belief as a relation to a sentence in mentalese. The intentional realist need not adopt a realist attitude toward RTM's individual theoretical representations to maintain that there is a fact of the matter about cognition. The further realist insistence is just not realistic given the general form and reach of cognitive psychological explanation. The functional relation to a sentence in mentalese representing a given attitude should not be regarded as disclosing the nature of the represented attitude at issue.

5 Conclusion

In a famous passage of *Philosophical Investigations*, Wittgenstein writes:

> And we may not advance any kind of theory. There must not be anything hypothetical in our considerations. We must do away with all *explanation*, and description alone must take its place. And this description gets its light, that is to say its purpose, from the philosophical problems. (2009: §109)

Such words have spawned the famous – many would say infamous – anti-theory legacy of the later Wittgenstein's philosophical oeuvre. How exactly to interpret them, on the other hand, is far from clear. It just isn't obvious what remains of philosophy once we do away with all theories and explanations. Aren't generalizations *about* philosophy, even a general claim to the effect that philosophical problems arise from "the bewitchment of our understanding by the resources of our language" (§109), theoretical contributions *within* philosophy? This much is clear, however: philosophy is predominantly theoretical. It advances theories. This is the uncontroversial backdrop to Wittgenstein's programmatic call, expressed with one 'may' and three 'must's, for philosophy's reorientation. But in the rush to replace philosophical explanation with "description alone", the Wittgensteinian impulse can easily overlook central features of the subject as the theoretical enterprise that it is.

Theoretical representations in the philosophy of language and mind can easily be mistaken for theoretical representations in the natural sciences that give rise to theoretical identifications à la gold being the element with atomic number 79. It can easily seem that when we theoretically represent things or facts pertaining to

semantic significance we are somehow purporting to lay bare what those things or facts really are at the end of the day. It can easily seem that when we theoretically represent mental states and episodes as involving mental representations we are somehow disclosing what those aspects of mentality really are at bottom. But this is a myopic view of how theoretical representations function within theories. Within the content program we model wide ranging phenomena by appealing to various theoretical representations that need not, and should not, be thought of as disclosing the nature of what they represent individually. When we theoretically represent a referential intention within a metasemantic theory as the intention to specify anything relevantly similar to samples in the environment of the speaker, we need not and should not assume that this representation itself reveals the nature of the attitudinal state in question. Rather, representing the intention figures within a broader explanatory effort to uncover the nature of aboutness for language and thought. When we represent a belief along Fodorian lines as the bearing of a relation to a sentence in mentalese, we need not and should not expect such a representation to disclose what it is for the represented belief state to obtain. Such cases are distorted by adopting a realist attitude towards the relevant theoretical representations, which in such light can seem fanciful. Instead of simply discarding those representations for not being sufficiently well grounded in the more experimental reaches of cognitive science, we can view them more holistically as playing larger roles within their respective theories that do not include disclosing the nature of whatever they represent individually.[16]

References

Donnellan, Keith. (1966) 'Reference and Definite Descriptions', *Philosophical Review* 75: 281-304.

Donnellan, Keith. (1970) 'Proper Names and Identifying Descriptions', *Synthese* 21: 335-358.

Fodor, Jerry. (1975) *The Language of Thought* (Cambridge, MA: Harvard UP).

Fodor, Jerry. (1987) *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (Cambridge, MA: MIT Press).

Fodor, Jerry. (1994) *The Elm and the Expert: Mentalese and Its Semantics* (Cambridge, MA: MIT Press).

[16]Many thanks to Roberta Ballarin, Jade Hadley, Graham Moore, and Ewan Townshend for discussions of this material. I gratefully acknowledge the support of the Social Sciences and Humanities Research Council of Canada.

Gòmez-Torrente, Mario. (2006) 'Rigidity and Essentiality', *Mind* 115: 227-259.

Häggqvist, Sören and Wikforss, Åsa. (2018) 'Natural Kinds and Natural Kind Terms: Myth and Reality', *British Journal for the Philosophy of Science* 69: 911-933.

Heim, Irene and Kratzer, Angelika (1998). *Semantics in Generative Grammar* (Malden, MA: Blackwell).

Kripke, Saul. (1979) 'A Puzzle About Belief', in Margalit, A. (ed.), *Meaning and Use* (Dordrecht: Reidel): 239-283.

Kripke, Saul. (1980) *Naming and Necessity* (Cambridge, MA: Harvard UP).

Lewis, David. (1975) 'Languages and Language', in Gunderson, K. (ed.), *Minnesota Studies in the Philosophy of Science* VII (Minneapolis: University of Minnesota Press): 3-35.

Lewis, David. (1979) 'Attitudes *De Dicto* and *De Se*', *Philosophical Review* 88: 513-543.

Marcus, Ruth B. (1990) 'Some Revisionary Proposals About Belief and Believing', *Philosophy and Phenomenological Research* 50: 133-153.

Needham, Paul. (2017) 'Determining Sameness of Substance', *British Journal for the Philosophy of Science* 68: 953-979.

Nolan, Daniel. (2006) 'Selfless Desires', *Philosophy and Phenomenological Research* 73: 665-679.

Putnam, Hilary. (1975) 'The Meaning of "Meaning"', *Minnesota Studies in the Philosophy of Science* 7: 215-271.

Quine, W.V. (1956) 'Quantifiers and Propositional Attitudes', *Journal of Philosophy* 53: 177-187.

Rescorla, Michael. (2020) 'Reifying Representations', in J. Smorthchkova, J., Schlicht, T., and Dolega, K. (eds.), *What Are Mental Representations* (Oxford: Oxford UP).

Simchen, Ori. (2004) 'On the Impossibility of Nonactual Epistemic Possibilities', *Journal of Philosophy* 101: 527-554.

Simchen, Ori. (2012) *Necessary Intentionality: A Study in the Metaphysics of Aboutness* (Oxford: Oxford UP).

Simchen, Ori. (2019) 'Realism and Instrumentalism in Philosophical Explanation', *Metaphysics* 2: 1-15.

Simchen, Ori. (forthcoming) 'Narrow Content and Parameter Proliferation', *Analytic Philosophy*.

Wittgenstein, Ludwig. (2009) *Philosophical Investigations*, 4th Edition (Oxford: Wiley-Blackwell).