Australasian Journal of Philosophy
Vol. 82, No. 1, pp. 199–212; March 2004

LEWIS ON INTENTIONALITY

Robert Stalnaker

David Lewis's account of intentionality is a version of what he calls 'global descriptivism'. The rough idea is that the correct interpretation of one's total theory is the one (among the *admissible* interpretations) that come closest to making it true. I give an exposition of this account, as I understand it, and try to bring out some of its consequences. I argue that there is a tension between Lewis's global descriptivism and his rejection of a linguistic account of the intentionality of thought. I distinguish some different senses in which Lewis's theory might permit, or be committed to, a kind of holism about intentional content, and I consider the sense in which Lewis's account might be said to be an *internalist* account, and the motivation for this kind of internalism.

David Lewis's account of intentional states and intentional content is an *internalist* one, in a sense: he is a proponent of narrow content. His strategy for solving the problem of intentionality is the most explicit and well developed internalist strategy that I know of, and I think it helps to bring out some consequences that any internalist account will have. Lewis's constructive proposals along with his criticisms of alternative strategies also throw light on a range of issues that keep recurring in the debates about intentionality, questions about the relation between language and thought, about holism, about the relation between internal and external perspectives on the content of speech acts and propositional attitudes.

Lewis's constructive proposals about intentionality were presented in a number of places, mostly in reaction to the views of others. His earliest discussion of these issues is in 'Radical Interpretation' [1974], which began as a response to Donald Davidson. In 'Reduction of Mind' in 1994, a general survey of his views about the mind, he characterizes his account of intentional content by contrasting it with the theory of a hypothetical philosopher that he named 'Strawman', a character who bears some resemblances to such real life philosophers as Michael Devitt, Jerry Fodor, and Hartry Field. The most detailed and direct response to the problem of intentionality was developed as a reply to Hilary Putnam's model-theoretic argument against a 'radically non-epistemic' conception of truth. First in 'New Work for a Theory of Universals' [1983], and then in more detail in 'Putnam's Paradox' in 1984, Lewis develops and defends a version of what he calls 'global descriptivism', and that will be the main focus of my discussion.

I will begin (in Section I) with an exposition of global descriptivism and with the twist that Lewis adds to the theory to avoid the anti-realist consequences that Putnam drew from a theory of this kind. In the remainder of the paper I will consider what Lewis's global descriptivism tells us about the relation between thinkers and the world they think about, and what the alternatives to it might be. In Section II, I will consider what difference it

makes whether one takes the problem of intentionality to be fundamentally a problem about language or a problem about thought. Lewis argues that the kind of account he contrasts with his own (Strawman's theory) gives an unjustified priority to language, but I will argue that Lewis's own account may be committed to the kind of linguistic picture of thought that he rejects. In Section III I will discuss holism. I will distinguish some different kinds of holism, arguing that while some are benign, others are more pernicious. While I agree that it is a virtue of Lewis's account that it permits a benign form of holistic representation, I will argue that his global descriptivism is also committed to holism of a more pernicious kind. In Section IV, I will conclude by considering what may be at issue between internalist and externalist accounts of intentionality. Lewis argues that it is a defect of the view he contrasts with his own that it fails to provide an account of narrow content, but I will question the need for narrow content, and argue that Lewis's account does not in any case provide a notion of narrow content that can play the role that such a notion is expected to play.

I. Global Descriptivism

In developing his positive account of intentionality Lewis framed the issue as a problem about the interpretation of a language. He adopted this strategy, not on principle, but simply because he was responding to a problem that Hilary Putnam had posed in linguistic terms. His response to Putnam began with this disclaimer:

> I shall acquiesce in Putnam's linguistic turn: I shall discuss the semantic interpretation of language rather than the assignment of content to attitudes, thus ignoring the possibility that the latter settles the former. It would be better, I think, to start with the attitudes and go on to language. But I think that that would relocate, rather than avoid, the problem; wherefore I may as well discuss it on Putnam's own terms.
>
> [Lewis 1999a: 57–8]

I think Lewis is right that a version of Putnam's problem will arise for a theory that begins with the attitudes rather than with their linguistic expression, but the relocated problem would take a different form, and it is not clear how Lewis's solution to Putnam's problem would translate into an account of the intentionality of thought that did not explain it in terms of the interpretation of a language. Since Lewis criticizes an alternative account of intentionality on the ground that it is committed to what he calls 'language-of-thoughtism', it is important to his project that he be able to state his own account in a way that does not acquiesce in the linguistic turn. I will consider this issue in section II below, but for now, let us follow Lewis in taking the problem to be a problem about how a language acquires its interpretation.

Lewis's story begins with a spirited defence of a description theory of names of the kind that Saul Kripke criticized in *Naming and Necessity*. According to descriptivism, the contribution that a name makes to the content of statements containing it is given by a description or cluster of descriptions of an individual. The referent of the term is what satisfies the description, or what best fits the cluster of descriptions. Lewis argues that if descriptivism recognizes a broader range of descriptions than is usually considered, it can be seen to have the resources to account for the phenomena that Kripke's took to refute

that account of reference. In fact, Lewis suggests, Kripke's examples and arguments point the way to a more adequate version of descriptivism. In particular, (1) descriptions may be rigidified (that is, the referent of a name in some counterfactual situation may be the thing that *in fact* satisfies the description in question), (2) descriptions might include information about tokens of the name itself, and its causal relations to things in the world (so the description associated with 'Socrates' might be something like 'the person who plays such-and-such a causal role in determining certain uses, by me and others, of the name "Socrates" '), and (3) the descriptions that define names might be interrelated, introducing names, not singly, but in families. So for example, the pair of names, $<$Jack, Jill$>$ might be defined by a cluster of descriptions that includes information about the relationship between the two.

But, Lewis emphasizes, even if a description theory of names and certain other terms can be defended, this is at best only a way of explaining the semantic properties of a part of the language in terms of other parts—in terms of the vocabulary used in the descriptions. To turn this kind of account into a general explanation of how language as a whole is interpreted, we need a *global* version of descriptivism, one that interprets all of the expressions of the language at once. The third of the points described above about the resources of the description theory (that names may be introduced in pairs, or families) points the way to such a global theory. Lewis suggests that we think of a whole language/theory as a cluster of descriptions that defines all of the descriptive vocabulary—predicates as well as singular terms—in the language. The language is interpreted relative to a *theory* stated in the language, where a theory is a consistent and deductively closed set of sentences. The correct interpretation of the language, according to global descriptivism, is the assignment of properties and relations to the predicates, and of individuals to the singular terms that comes as close as possible to making the theory (or most of its claims) come out true. What the theory being interpreted says, on this account, will be that the world is such as to provide an interpretation of the language that is a model for the theory. The content of the theory as a whole could be stated explicitly by replacing all non-logical expressions of the theory by variables of the appropriate type and prefixing the whole theory by first and second order existential quantifiers binding all of these variables.[1] So the content of the theory is that there exist properties, relations, and individuals that relate to each other in the way that the sentences of the theory specify.

As Lewis made clear, without some further constraints on interpretation, it will be far too easy to find models, whatever the world is like, for any consistent theory. This means that unconstrained global descriptivism fails to provide any explanation for how a language can make any substantive claims at all about the world. That is the upshot of Putnam's model-theoretic argument, or as Lewis calls it, Putnam's paradox. Lewis's response is not to abandon global descriptivism, but to look for additional constraints to impose on the admissible interpretations, and what he proposes is a constraint on the domains of properties and relations that are allowed to be values of the variables that replace the predicates in the theory. Before looking at this proposed further constraint, and

[1] That is, the theory is replaced by its *Ramsey sentence*. This method of eliminating theoretical terms is used in several other places by Lewis. See, for example, Lewis [1972].

at Lewis's response to Putnam's argument that no further constraint will work, let me try to say in a little more detail how the interpretation procedure is supposed to work.

Lewis's global descriptivist account of interpretation is like the liberal cluster version of the local descriptive theory in that it allows for only approximate fit, and this is essential to the viability of the theory. Just as (according to the local cluster theory) a name may succeed in referring even if some of the descriptions in a cluster of descriptions that defines the name fail to be satisfied by it, so (according to global descriptivism) the world may succeed in interpreting the predicates and terms of a theory even if some of the claims of the theory are false, on that interpretation. A strict version of global descriptivism would be untenable since it would have the consequence that either the theory as a whole (which is intended to represent the totality of a speaker's beliefs) is true (which means that the speaker has only true beliefs), or else all of the predicates fail to have extensions in the actual world (which implies that the speaker has no true beliefs). This can't be right if, as seems intuitively plausible, most of us have some true and some false beliefs.

But the point cannot be that all the theory asserts is the approximate truth of its claims. To make sense of the possibility of some true and some false beliefs, we need to separate the role of a theory T in interpreting the vocabulary from its role in stating the content of the speaker's beliefs. The rule of interpretation should be something like this: *The extensions, in possible world x, of the predicates and terms of the language are those given by the interpretation that does the best job of making theory T approximately true*. But once the predicates get this interpretation, then the individual claims of the theory, as expressions of the speaker's beliefs, are interpreted strictly.

So what, according to global descriptivism, does a particular statement of the language assert? For example, suppose the language has a one-place predicate F and a name a. The sentence Fa will be true in possible world x if and only if the individual that the interpretation procedure assigns to the name a (relative to world x) has, in world x, the property that the interpretation procedure assigns to predicate F (relative to world x). The interpretation procedure may assign different individuals and properties to the name and the predicate relative to different possible worlds, so the propositional content of the sentence will not be that a certain individual has a certain property, but rather that whatever individual fits the global description associated with a has whatever property it is that fits the global description associated with F. Since the whole theory T is involved in the interpretation of each name and predicate, even the simplest statements of the language will contain an implicit reference to the total theory.

This point holds even if the descriptions that are implicit in the meanings of names and other terms are rigidified descriptions. For example, even if the description that the theory associates with the name 'Aristotle' is something like 'the *actual* person who was the last great philosopher of antiquity', it will still be true that the (narrow) content of our statement, 'Aristotle was fond of dogs' will be a proposition that is true in world x, if and only if the person who was the last great philosopher of antiquity in world x was fond of dogs in world x (or more strictly, stood in whatever relation plays, in world x, the fondness role to whatever things play the dog role in x.) The only role of rigidification, on the global descriptivist account, is to isolate descriptions from the effects of modal operators. So the content of 'It is possible that Aristotle was fond of dogs' would be something like this: it is true of whoever was the last great philosopher of antiquity that possibly he was fond of dogs. (And if the descriptive content of 'dog' is rigidified as well, then the content will be

something like this: it is true of whoever was the last great philosopher of antiquity and whatever kind it is that plays the 'dog' role that it is possible that the former was fond of the latter. We should probably do the same for 'fond of', but you get the idea.)

But as we have noted, global descriptivism would be a nonstarter if it were not supplemented with some additional constraints. Let us say that the language must be interpreted, not by just any model, but by an *admissible* model for the theory; the task of filling out a tenable version of global descriptivism is the task of specifying the constraints on the class of models that are admissible. One might require that an admissible interpretation be one that meets a *causal* constraint: the values of the primitive expressions of the language must be those that play a certain role in the causal explanation for the fact that the speaker uses those expressions as she does. That was Michael Devitt's proposal [1983], but Lewis preferred to locate whatever causal condition plays a role in determining reference in the content of the descriptions, rather than in an external constraint on the interpretation of terms. He proposed instead a *metaphysical* constraint: the values of the predicate variables of the language must be more or less *natural* properties and relations. The model-theoretic argument that any consistent theory will have a model (an interpretation of the descriptive terms of the language relative to which the theory is true) depends on allowing any set of n-tuples of individuals to be the value of an n-place predicate. If interpretations are restricted by ruling out excessively artificial, gruish, gerrymandered values for the predicates, then it may be that the best admissible interpretation of a theory will be one that makes it only approximately true, or it may happen that there will be no admissible interpretation for a theory, in a given possible world.

Lewis's account of interpretation requires a metaphysical commitment but it is a very abstract one, and while it is controversial, it is independently motivated, and I think difficult plausibly to deny. What the account is committed to is only the rejection of an extreme form of nominalism that denies that there is any distinction to be drawn (independently of our theorizing) between more natural and more artificial ways of grouping individuals into classes. No particular account of universals is presupposed by Lewis's account, and nothing is assumed about the particular character of the properties and relations that are the more natural ones (relative to the actual world).

Putnam expressed scepticism that the world as it is in itself can supply the distinctions between properties and kinds that give us the joints at which our words and concepts carve it, but his main reason for thinking that a strategy like Lewis's cannot avoid his paradoxical conclusion is that he thinks that any proposed constraint will face the 'just more theory' objection, which goes like this: Any constraint that we might propose will be a bit of theory that we propose to assert, and so will itself require interpretation. It should therefore be added to our total theory. But then the model-theoretic argument applies to the constraint itself. So long as the original theory, augmented with a statement of the constraint, is consistent, it will always be possible to find an interpretation relative to which the theory, including the constraint, is true. If the constraint involves a distinction between natural and artificial properties and relations, along with some theoretical postulates spelling out what the distinction comes to and a principle that requires all values of predicates to be natural, then we will be able to find an interpretation for the predicate 'natural' that makes the postulates true, and makes the values of all predicates ones that satisfy 'natural' on that interpretation. Whatever the character of the further constraint, Putnam argues, it can be shown to be ineffectual in this way.

Lewis's response to the 'just more theory' objection is simple, and it is a response that is perfectly general, applying to any constraint *C* that might be proposed. 'The constraint', Lewis writes, 'is *not* that an intended interpretation must somehow make our account of *C* come true. The constraint is that an intended interpretation must conform to *C* itself' [Lewis 1999a: 62].[2] The point is that it is not the subject being interpreted who is supposed to affirm the proposed constraint—to add it to his set of beliefs. Rather, the constraint is affirmed by the theorist giving an account of what it is that determines the content of the sentences that the subject accepts. The constraint is part of an external account of how speakers, viewed as things in the world, are related to the things they talk about.

The critic may be suspicious of this response. The theorist who is proposing the constraint is also a speaker/thinker, so the constraint, in his mouth, is still a piece of theory in need of interpretation. One might try to explain how the theorist achieves determinate reference by postulating another external constraint (or a version of the same constraint), in the mouth of another theorist, but doesn't this lead to a vicious regress? Lewis's answer is no, and I agree. The response does point to a way in which this account of intentionality is an externalist one, but that is not a reason to think that it is not successful. It is a general feature of any theory about how language relates to the world that it must be stated in a language, and the theory will be able to say how language can be about the world only if the theorist's language itself succeeds in being about the world. If an explanation of intentionality is to be adequate, it had better be that the account makes sense of the fact that among the things we are able to talk about are ourselves and our relations to the things we talk about. But as Lewis argues, our task is not to refute, on his own terms, the sceptic who doubts that we are able to talk about the world at all.

I find Lewis's response to the 'just more theory' objection wholly persuasive, but I remain sceptical about whether his positive account of intentionality gives a plausible account of the contents of our thought and speech. In the remainder of the paper, I will point to some problems that I think it faces.

II. Language and Thought

As noted above, since Lewis's account of intentionality is developed in the context of a criticism of Putnam's model-theoretic argument, he follows Putnam in posing the problem of intentionality as a problem about the interpretation of language. But he makes clear that he would prefer to begin with thought, and to explain the intentionality of language as derivative from the beliefs and intentions of language users. Lewis criticizes his foil, Strawman, for assuming that mental representations have a linguistic structure. 'I don't believe that folk psychology says that there is a language of thought. Rather, I think it is agnostic about how mental representation works—and wisely so' [Lewis 1999b: 310]. I agree with Lewis here, but my worry is that I don't see how his positive account of the way a language is to be interpreted can be translated into an explanation of the intentionality of thought that does not itself presuppose the linguistic picture of thought that he rejects.

[2] Lewis cites Devitt [1983], who makes this point in defence of a causal constraint.

More generally, I suspect that posing the problem in terms of language has a distorting effect on our understanding of both thought and language.

Before getting to the main issue, let me just point to one place where unclarity about the relation between language and thought obscures an issue. As Lewis notes, one of the sources of Putnam's paradox is a voluntaristic view of reference—the assumption that, in Putnam's words, '*we* interpret our language or nothing does' [Putnam 1980: 482]. Lewis takes the main lesson of Putnam's paradox to be the untenability of the idea that reference is determined by the speaker's referential intentions [Lewis 1999a: 63]. But we should not need Putnam's paradoxical argument to see the blatant incoherence in the very idea of combining the thesis that linguistic reference is to be explained prior to and independently of thought with the thesis that linguistic reference is to be explained in terms of the referential intentions of the speaker. Once we reformulate the problem of intentionality as a problem about beliefs and intentions, and see the voluntaristic view of reference and meaning as a view about how the beliefs and intentions of language users determine the representational properties of the language they use, then it can be seen that neither Putnam's model-theoretic argument nor externalist accounts of mental and linguistic content present obstacles to a voluntaristic (or at least intentionalist) theory of *linguistic* reference and meaning.

Our main problem is to see how Lewis's positive account of interpretation, developed in the context of a discussion of Putnam, is to be reconciled with his preference for an account of intentionality that begins with propositional attitudes, rather than with language, and that does not presuppose a language-of-thought conception of the attitudes according to which whenever it is true that x believes that P, there is some quasi-linguistic token in x's brain that says that P. The problem of intentionality, in this context, is the problem of explaining what it is about the behavioural capacities and functional organization of the subject and the ways it is disposed to interact its environment, that constitutes the fact that it has beliefs and other intentional mental states with certain contents. Lewis takes the familiar dispositional account of belief and desire to be implicit in folk psychology, and to provide an answer to the question.

> A system of beliefs and desires tends to cause behaviour that serves the subject's desires according to his beliefs. ... Beliefs change constantly under the impact of perceptual evidence: we keep picking up new beliefs, mostly true, about our perceptual surroundings; whereupon our other beliefs (and our instrumental desires) change to cohere with these new beliefs.
>
> [Lewis 1999b: 320]

Lewis refers to the conception of rationality that is implicit in folk psychology (and that is refined in decision theory) as '*constitutive* rationality'. The intentional mental states of an agent have the content that they have in virtue of the fact that the behaviour of the agent can be explained as rational behaviour on the hypothesis that the agent's mental states have that content.

This strategy for explaining the basis of attribution of content is agnostic about the medium of mental representation because it founds the attribution of content, not on the vehicle of representation, but on the consequences that states with content have for action. One does not start by identifying a candidate to be a representation (a set of mental sentences, or maps, or whatever), and then ask how the content of that vehicle of

representation is determined. For all folk psychology says (according to this kind of account) the mechanisms that realize the dispositional properties that constitute beliefs with a certain content might have diverse distributed components that cannot, individually, be interpreted as representations at all.

One can see how an analogue of Putnam's argument—at least an argument with an analogous conclusion—will present a problem for a pure dispositional account of intentional states. The problem, in this context, is familiar: Just as it is too easy to find an interpretation that makes any consistent theory true, so it is too easy to find attributions of beliefs and desires to an agent relative to which the agent's behaviour is rational in the sense that it would tend to realize the agent's desires in a world in which his or her beliefs are true. Just about any beliefs can be rationalized by adjusting the attribution of desires, and vice versa. And the response to this problem also requires a move that parallels the response to Putnam's problem: we need further external constraints on interpretation if we are to avoid pervasive indeterminacy.

Some of the things Lewis says suggest that further constraints are already built into the idea of constitutive rationality. Among the things that, according to Lewis, folk psychology says are that our beliefs are mostly true, that perceptual beliefs are about the believer's perceptual surroundings, and that they change under the impact of perceptual evidence. If constitutive rationality explains content, not wholly in terms of the behaviour that intentional states dispose the agent to engage in, but in part in terms of the way those states are caused, then it may have the resources to avoid the kind of pervasive indeterminacy that threatens the tenability of the account. But in developing his response to Putnam's paradox, and in his criticisms of Strawman, Lewis seems to reject the idea of building causal constraints into the conditions determining intentional content. He is a defender of causal descriptivism, rather than a causal theory of reference. The worry is that if causal conditions are built into the characterization of the way content is determined, we will not get an account of narrow content.

Lewis's metaphysical constraint was supposed to provide a constraint that avoids the pervasive indeterminacy, as he acknowledges that a causal constraint would, but that also yields a notion of content that is narrow in the sense that it is determined by intrinsic properties of the subject, as a causal constraint would not. But how does the metaphysical constraint apply in the constitutive rationality context? If we are given an uninterpreted language, with a well defined syntax, and a theory stated in it, as the material to be interpreted, then it will be reasonably clear how Lewis's constraint works. An interpretation matches properties and relations to the predicates in a way that provides an interpretation relative to which the theory is true. An *admissible* interpretation will be one that provides a model for the theory, and that also meets the condition that the properties and relations that are the values of the predicates are drawn from some restricted set. But where our task is to interpret the dispositions that explain an agent's behaviour, what does the metaphysical constraint—the restriction to more or less natural properties and relations—constrain? The metaphysical constraint works on the internal structure of a given theory, constraining the interpretation of its primitive constituents. If what is given is a pattern of behaviour rather than a language with predicates to be interpreted, it is not clear how it is to be applied.

What we need is a constraint on global belief states, rather than on theories that might be used to express the content of a global belief state. That is, we need a constraint on the

classes of possible worlds that are candidates to be the class of worlds compatible with someone's beliefs. One might try to use Lewis's metaphysical constraint on the interpretation of theories to constrain global belief states (represented by classes of possible worlds) in something like the following way: First, note that every consistent theory determines a class of possible worlds, by Lewis's procedure: those worlds that provide an admissible model for that theory. Say that a class of possible worlds is a *candidate belief state* only if it is a class determined, in this way, by some theory, and require that the dispositional account of belief choose among only belief states that are candidates in this sense. But a constraint of this kind would be considerably weaker than Lewis's, and it is not clear that it would be responsive to the indeterminacy problem, as it arises in this context. The argument that rationality of just about any behaviour can be reconciled with just about any attributions of belief by adjusting the attributions of desire does not seem to depend on an appeal to attributions of belief that would violate this constraint.

I conclude that Lewis's strategy for avoiding indeterminacy in the attribution of attitudes while retaining an internalist conception of intentionality is more closely tied to a linguistic picture of mental representation than he thought. There is a tension between his account of constitutive rationality and his response to Putnam's paradox, a tension that I think is best resolved by recognizing that constitutive rationality is an externalist theory about the relation between rational agents and their environment, and that the thoughts of such agents do not have a kind of intentional content that is wholly determined by their intrinsic properties. But criticism of the internalist project and defence of a causal constraint on intentional content should not be taken as a defence of Lewis's foil, Strawman. Strawman proposed a causal constraint on reference, in the context of a linguistic account of intentionality. His causal theory of reference is atomistic, aiming to explain the relation between expressions and what they refer to independently of the role of the expressions in the general account of constitutive rationality. I share Lewis's doubts that a defensible causal theory of reference of this kind (not just for names, but for non-logical expressions of all kinds) can be constructed, but I think Strawman's mistake is not that he finds a causal dimension in reference; rather it is that he does not recognize that the causal dimension of reference has its source in more holistic causal constraints on beliefs and intentions, together with an explanation of reference in terms of beliefs and intentions.

III. Holism

Lewis is critical of Strawman for being committed to an atomistic conception of representation.

> A serious issue, and one on which I take folk psychology to be agnostic, concerns the relation between the whole and the parts of a representation. Suppose I have a piece of paper according to which, *inter alia* Collingwood is east of Fitzroy. Can I tear the paper up so that I get one snippet that has exactly the content that Collingwood is east of Fitzroy, nothing more and nothing less? If the paper is covered with writing, maybe I can; for maybe 'Collingwood is east of Fitzroy' is one of the sentences written there. But if the paper is a map, any snippet according to which Collingwood is east of Fitzroy will be a snippet according to which more is true besides.
>
> [Lewis 1999b: 310]

Lewis suggests that if the way information is represented is holistic—map-like, or hologram-like, rather than sentence-like—then plural propositional attitude terms such as 'beliefs' are 'bogus plurals': 'You have beliefs the way you have the blues, or the mumps, or the shivers' [Lewis 1999b: 311]. But this is not right, since whatever the nature of the vehicle or vehicles of mental representation, the plural noun 'beliefs' does not refer to that vehicle, or to those vehicles. What it refers to is the *contents* of a representation—to the propositions that are believed. And even if there is a single map-like internal structure in a believer that make it the case that she has the beliefs that she has, that structure will determine a plurality of propositions that are believed. I agree with Lewis that folk psychology is agnostic about the medium of mental representation, and that our account of propositional attitudes should allow for the possibility that the contents of attitudes are determined by representations of diverse forms. But whatever the character of what is going on in a person's head when she has beliefs, those goings on should not be confused with what the person believes.

Holism, in Lewis's snippet sense, is a thesis about the form of mental representation, and is an issue on which folk psychology should remain noncommittal. But I think Lewis's reason for insisting on agnosticism about the forms of representation is that he is committed to holism in a different sense that concerns the character of the problem of intentionality—the problem of saying what it is about speaker/thinkers and their place in the world that explains their capacity to represent, and that gives their representations the content that they have. According to the proponent of this different holistic thesis, one cannot solve the problem by starting with the atoms of representation, and then explaining how more complex representational structures get their representational properties in terms of the representational properties of their simple parts. Instead, one must explain representation first in terms of larger units, and then explain the representational properties of simpler parts in terms of their role in the larger structure. The real problem with Strawman's account, I think Lewis would say (and here I would agree), is not that he goes out on a limb about the medium of mental representation, but that he is pursuing an unpromising strategy for explaining intentionality. The constitutive rationality strategy is holistic in this second sense, as is Lewis's strategy of global descriptivism. Here the philosopher does not remain agnostic. Even if the medium of mental representation turns out to be linguistic, and can be cut up into snippets each of which expresses a proposition, one should still expect the explanation of what gives those snippets the content that they have to be holistic.

Holism has been thought by some philosophers to be a pernicious doctrine, incompatible with stable explanations of behaviour in terms of the beliefs, desires, and values of a rational agent. Jerry Fodor, for example, wrote 'Meaning Holism looks to be entirely destructive of the hope for a propositional attitude psychology' [1987: 56]. But I think the kind of holism that Fodor is worried about is different from either of the two kinds that I have distinguished, so let me point to a third holistic doctrine that might be thought to apply to a representational structure.

As we have noted, whatever form mental representation takes, any total state of belief, or more generally, any representation (a map, a chart, a set of sentences, or whatever) will determine a set of propositions—the propositions that are believed by one who is in that state of belief, or that must be true for the representation to be accurate. Now suppose that we have such a representational structure, and that a small change occurs in it. If it is a set of

sentences, remove just a few snippets and replace them with others. If it is a map, perhaps just interchange Fitzroy and Collingwood. According to the version of holism that I am now trying to characterize, a representation is holistic if any such small change will inevitably bring about a massive change in the set of propositions that the representation determines.

This kind of holism is independent of whether the medium of representation is holistic in Lewis's snippet sense. The map is holistic in the snippet sense, so there will be no nice match up between the individual propositions represented and bits and pieces of the map, but that does not imply that the map will be holistic in our third sense. The small change in the map will mean that the proposition that Collingwood is east of Fitzroy will be dropped from the list of propositions represented, along with some others, but most of the propositions that would be naturally expressed in saying what the map represents will remain the same. On the other hand, a representation may be atomistic in Lewis's snippet sense, while being holistic in our third sense. Even if the representation is a text whose sentences correspond one-one with the items on a list of propositions represented, it might be that a small change in the text will force a massive change in the corresponding list of propositions, since the change might affect the interpretation of all the sentences in the text that remains.

I don't think the broad constitutive rationality strategy for explaining intentionality is committed to this third kind of holism—meaning holism—which I agree is pernicious, but I think it is clear that Lewis's global descriptivism is committed to it. Suppose that a person's total state of belief can be represented by a theory and that it is interpreted in accordance with global descriptivism, supplemented with Lewis's metaphysical constraint on the admissible models. The set of propositions believed will correspond to the theorems of the theory, but as we noted above, for the global descriptivist, the theory as a whole is playing two roles: determining the sentences that express the beliefs, and providing an interpretation for those sentences. The whole theory enters into the interpretation of each sentence: a sentence of the form *Fa* says that the individual, whatever it is, that best fits the global description that the total theory associates with *a* has the property, whatever it is, that best fits the global description associated with *F*. A small change in the theory may leave the sentence *Fa*, and most of the other sentences, on the list of sentences that express beliefs, but both the predicate *F* and the descriptive name *a* will change their meanings, and the sentence along with most of the others, will express different propositions.

My characterization of meaning holism is vague, since I am providing no measure on the size of a change, either in a representational structure, or in the class of propositions that it determines. For any two total theories with different content, there will be infinitely many propositions (sets of possible worlds) that are entailed by both theories, and infinitely many that are entailed by one, but not the other. But if we restrict ourselves to propositions that are expressible in the theories in question, then the holism thesis will hold, and this is enough to get pernicious consequences from the thesis. It implies that if two speakers/thinkers disagree about anything, than they will disagree massively about the things that either of them can say, and that if a person's beliefs change at all, then they change massively, even if the person will use many of the same sentences to express his or her new beliefs. This makes it difficult to give a coherent account of communication, or of belief change, in terms of the contents of speech acts and propositional attitudes.

The global descriptivist may concede that the narrow content of an agent's statements and attitudes may be idiosyncratic and unstable in this way, but argue that this is not so serious, since the statements of a theory may also have wide contents which derive from the

rigidification operators that (he hypothesizes) are used liberally in the global theory that gives the best account of the semantics of our actual language and thought, and that the wide contents will be more stable. For example, he may grant that the content of the description that my total theory associates with the name 'Aristotle' is different from the description that plays a corresponding role in your total theory, and so the contents of all of our statements about Aristotle, using that name, will be different. But if both descriptions are rigidified, and if they both, in fact, determine the same referent, then our statements using the name 'Aristotle' will have the same truth value, even when 'Aristotle' occurs in a modal context. I may not know or understand exactly what you meant, but that won't matter so long as the world cooperates, ensuring that we succeed in referring to the same thing. But if we explain communication of information and the way in which attitudes evolve in response to new information in terms of a kind of content that is determined in part by the external environment, why should we go to such lengths to give an internalist account of intentionality in the first place?

IV. Internalism and Externalism

There are many dimensions to the contrast between an internalist and an externalist picture of intentionality, and one may have different motivations for seeking an explanation of intentionality wholly in terms of the intrinsic properties of a speaker/thinker. In what ways, and for what reasons, was Lewis an internalist? Some internalists may be motivated by a representationalist picture that aims to explain content attribution in terms of linguistic or quasi-linguistic representations. (The very term 'content' may encourage the internalist picture. It may suggest that the content of a representation is something you will find within it. The representation is the vehicle, and the content is riding inside.) But this was not Lewis's reason. He would agree that when we say something of the form '*x* believes that *P*', we are not referring, with the 'that *P*', to any kind of internal representation, either one that is language-like, or one that is more holistic, like a map. We are referring to a proposition (or perhaps to a property) and using it to classify a person's state of mind, or to explain the role of a linguistic act in a communicative practice. Lewis may have been a fan of narrow content, but he was clear that meanings and contents (whether narrow or wide) ain't in the head: There are not (Lewis would agree) two kinds of content, narrow and wide, but two roles that a single kind of abstract object may play in the characterization of a state of mind. What is narrow or wide is the property of having an attitude with a certain content, and the contents themselves are objects that are explained in terms of the way an (external) world might be.

Some may be attracted to an internalist strategy for explaining intentionality because they seek to explain from within how we are able to reach out and (mentally) touch the world, but this was not part of Lewis's motive for seeking a notion of narrow content. His project was explicitly externalist in that it saw the problem, not as the problem of building a conception of the world from within, but as the problem of situating a speaker/thinker in the world that we find ourselves in, and explaining how the facts about the person, the world, and the possible worlds determine the contents of his or her thoughts. Lewis emphasizes that his project is externalist in this sense in his response to Putnam's 'just more theory' objection.

One of Lewis's motivations for his internalism was his pessimism about the prospects for a particular kind of externalist theory—Strawman's theory. He thought that there was no hope of success for an atomistic causal theory of reference, a theory that tries to explain intentional content in terms of the way primitive expressions of a language of thought are causally connected with the things they denote. But this is a reason that applies only to this particular externalist strategy for explaining intentionality. Lewis also had more general objections to Strawman's externalism: he thought that it was clear, intuitively, that our intrinsic twins who inhabit various counterfactual worlds (Swampman, the brain in the vat, or Oscar on Twin Earth) have beliefs in common with us. But I think a thoroughgoing externalist can account for our intentional similarities with these counterfactual creatures without admitting any notion of content that is determined wholly by the intrinsic properties of the believers. In some cases (Oscar on Twin Earth) we explain the common beliefs in terms of common features of our different environments. In others (Swampman, for example) we may treat them, by courtesy, as members of our kind and community because of their (contrived) similarity to us. To the extent that it seems intuitively right to attribute beliefs to Swampman, it is beliefs with wide content that we attribute. It is water that he wants to drink, if he is thirsty, and that he believes is to be found in the lake.

Lewis should agree that we can attribute beliefs with content to Swampman only if we treat him as a member of our kind, and so allow for the possibility that the contents of his attitudes will depend on the way ordinary human beings are situated in the world. Lewis's version of functionalism is not individualistic, and so it is not clear that his account of constitutive rationality, even in its pure dispositional form, delivers a notion of narrow content that will satisfy an internalist who believes that mental states supervene on the intrinsic properties of individuals.

> Narrow content is independent of what you are acquainted with, but that does not mean that it is altogether intrinsic to you. For it still depends on the causal roles of your brain states; . . . it is the *typical* causal role of your brain states that matters. But you may be an atypical member of your kind; hence what is typical of your kind is not intrinsic to you. So I can only say this: if X and Y are intrinsic duplicates, and if they live under the same laws of nature, and if they are of the same kind, then they must be exactly alike in narrow content.
>
> [Lewis 1999b: 315]

Once it is granted that intrinsic duplicates may be of different kinds, and therefore differ with respect to the contents of their beliefs, it is no longer clear what motivates the search for a kind of content that is narrow in this qualified sense. It seems reasonable to expect that the theory of constitutive rationality will need to be construed broadly, taking account of the ways thinkers interact with their environment in order to give a plausible account of content, and that once one gives up the bad reasons for being an internalist, there is little reason to work to find a theory that avoids this construal.

I have argued that there are some tensions between two different parts of Lewis's account of intentionality—his global descriptivism and his conception of constitutive rationality. The first, if it worked, might provide a notion of narrow content, but it is hard to reconcile with Lewis's rejection of the language-of-thought picture, and it seems to commit him to a kind of holism that is intuitively unacceptable. The constitutive rationality strategy for explaining intentionality does not yield any notion of propositional content

that is determined wholly by the intrinsic properties of the thinker, but it does not need one. If one is clear that the facts about the world in which a speaker/thinker finds himself play a role in determining the content of his speech acts and thoughts, not by way of an atomistic causal theory of reference for a language, but through causal constraints that are part of a theory of constitutive rationality, then Lewis's reasons for resisting an external causal constraint on the determination of content may not have the force that he took them to have.

Massachusetts Institute of Technology

REFERENCES

Davidson, D. 1973. Radical Interpretation, *Dialectica* 27: 313–28.
Devitt, M. 1983. Realism and the Renegade Putnam: A Critical Study of *Meaning and the Moral Sciences, Noûs* 17: 291–301.
Fodor, J. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge MA: Bradford Books/MIT Press.
Lewis, D. 1972. Psychophysical and Theoretical Identifications, *Australasian Journal of Philosophy* 50: 249–58.
Lewis, D. 1974. Radical Interpretation, *Synthese* 23: 331–44.
Lewis, D. 1983. New Work for a Theory of Universals, *Australasian Journal of Philosophy* 61: 343–77.
Lewis, D. 1999a (1984). Putnam's Paradox, in *Papers in Metaphysics and Epistemology*, Cambridge: Cambridge University Press: 56–77.
Lewis, D. 1999b (1994). Reduction of Mind, in *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press: 291–324.
Putnam, H. 1980. Models and Reality, *Journal of Symbolic Logic* 45: 464–82.