***The Metaphysics of Representation***, by J. Robert G. Williams. Oxford: Oxford University Press, 2020. Pp. xxii + 213.

## 1. Introduction

In this superb book, Williams sets a very ambitious goal for himself: to sketch biconditionals that define representational conditions in non-representational terms (p. xvii). Representation is not a spooky, primitive capacity of the mind; it is built from more basic ingredients. At the center is his radical interpretation theory of belief and desire, inspired by the work of David Lewis. To a first approximation:

> **Basic radical interpretation theory**. The correct assignment of beliefs and desires to an agent is the *most rationalizing assignment* given her *perceptual evidence* and dispositions to *act*. (p. 16, 97ff)

Williams does not give a master argument for this account of belief and desire over rival accounts in which constitutive rationality plays no role. Rather, his main goal is the laudatory one of *theory-building*. In this respect, his book hearkens back to the decade or so from the early 80s to the early 90s that was the heyday for developing grand theories of representation. In particular, his main aim is *to develop the details* of the basic radical interpretation theory – something Lewis never fully did. The result is a unique, multi-stage theory of representation that importantly departs from Lewis in many places.

Sections 2 and 3 of this review focuses on Williams' development of his radical interpretation theory of belief and desire. Section 4 takes a look at his innovative theory of linguistic representation.

## 2. Williams' Radical Interpretation Theory

Throughout his book, Williams focuses on a hypothetical person, Sally. To keep things simple, let's at first assume that Sally doesn't yet speak a language; she is one of our pre-linguistic ancestors.

On Williams' radical interpretation theory, how do the non-representational facts about Sally fix her beliefs and desires?

Williams' radical interpretation theory offers a reductive account of Prelinguistic Sally's beliefs and desires in terms of the *most rationalizing interpretation* given her *perceptual evidence* and dispositions to *act*. On pain of circularity, in explaining these things, he cannot appeal to facts about Sally's beliefs and desires. In fact, to achieve his reductive aim, he cannot appeal to

any representational facts about Sally at all. Here is how Williams proposes to fill in these three slots of the theory.

(i) *Sally's perceptual evidence.* Williams holds that Prelinguistic Sally's perceptual evidence consists in representational facts of the form: *I'm having a perception that represents that something is F* (p. 179ff). In turn, appealing to Karen Neander's reductive externalist-teleological theory of perceptual representation, he reduces these representational facts that make up Sally's perceptual evidence to facts of the form: *I'm in an internal physical state S such that S is a sensory-perceptual state and S has the systemic function to be produced in response to Fs* (p. 185ff). So Sally's reasons for her beliefs about the world ultimately derive from the *wide*, *externally-determined* contents of her internal sensory-perceptual states.

(ii) *Sally's actions*. How do we get from Sally as a physical system to Sally as performing-various actions, which are candidates for rationalization in terms of belief and desire? Williams argues that the objects of rationalization are actually 'proto-decisions' (p. 174ff). For example, Sally might form a proto-decision *to move away from a flying projectile*. This proto-decision represents some basic behavior to be performed. To explain this, Williams adapts Neander's theory: one's intentional-motor system outputs an internal state that has the systematic function to *produce* the relevant bodily movement (p. 188ff).

(iii) *The most rationalizing interpretation*. Once we have figured out Prelinguistic Sally's perceptual evidence (facts about what her inner states have the function of detecting) and her dispositions to proto-decide, we can work out the *most rationalizing belief-desire interpretation* of her (97ff). Williams argues that, to rule out deviant interpretations, the most rationalizing interpretation must be one that maximizes 'substantiative rationality' as well as mere 'structural rationality'. His argument – which builds on some remarks of Lewis – concerns what he calls the 'bubble puzzle' (for details see p. 17ff). Roughly, structural rationality is a matter of consistency of belief, means-ends coherence, and so on. And 'substantive rationality is a matter of reason-responsiveness, where there may be epistemic reasons for belief, or practical reasons for action' (p. 26). The point is not that Prelinguistic Sally cannot have irrational beliefs and desires; rather, the correct interpretation will minimize irrationality. Although Williams' reductive account of representation appeals to normative facts, he does not develop a naturalistic reduction of such facts (p. 13).

That, then, is Williams' basic theory. To see how it might work, consider an example involving Prelinguistic Sally. A nasty person from an enemy tribe hurls a rock toward Sally, and she moves out of its path in the nick of time. She believes that it is headed toward her, and she desires that it not hit her and cause her pain. But there are many perverse interpretations. One of them is that she believes that the rock is moving *away*, she desires that it instead fly

toward her and hit her and cause her pain, and she believes that by moving away she will magically cause the rock to reverse direction and hit her. What determines that the first interpretation is correct and the second is incorrect? According to Williams, the correct solution to such underdetermination worries appeals to facts about rationality. Sally has a reason to believe that a rock is headed toward her, a reason to desire to avoid being hit on the head and feeling pain, and a reason to believe that if the rock is moving toward her then she can avoid being hit by moving away. So the first interpretation maximizes her rationality, while the second gratuitously attributes gross irrationality to her. That is what singles out the first interpretation as the *correct* interpretation. Further, Williams has a story about the sources of Sally's reasons. For instance, Sally has a reason to believe that the rock is moving toward her, because the perceptual evidence she conditionalizes upon includes the fact that she is in an internal state that has the historical function of being caused by *something moving toward her* in the external world.

Perhaps Williams' theory can deliver plausible verdicts about Prelinguistic Sally's other beliefs and desires, including false and irrational beliefs, beliefs and desires about the unobserved, and so on. It accommodates the intuition that there is *some* kind of constitutive connection between Prelinguistic Sally's beliefs and desires and her dispositions, while at the same time avoiding a crude behaviorism.

Williams often notes that radical interpretation theory is schematic. He develops one way of filling in the slots. But he is open to alternatives (p. 202).

In particular, others who favor the radical interpretation approach might opt for an alternative, more internalist and 'consciousness-first' account of the source of Sally's reasons for her beliefs. It is plausible that her experiences have 'phenomenal contents' fixed by phenomenology, and that phenomenology is narrow. Given these two claims, it follows that, in addition to having 'wide' contents determined by connections to the environment, Sally's experiences have narrow phenomenal contents. For instance, in the above example, in addition to having the function to be caused by a thing moving toward her, Sally's experience has a narrow 'phenomenal content' *that a thing is moving toward me*. The narrow phenomenal content, but not the wide content, is bound up with how things phenomenally appear to her. Williams' externalist-teleological theory of perceptual representation does not apply to such internally-determined experiential representation. (Perhaps some other internalist reductive account does.) It is further plausible that it is just in the nature of having an experience with a certain phenomenal content that, if you have this experience, you have a reason to believe that content (Pryor 2000, fn.37). Putting all this together, Prelinguistic Sally's reasons for her beliefs about the external world derive, at least in part, from her having conscious experiences

with certain narrow contents. Indeed, some hold that it is only individuals with conscious experiences that have any reasons for belief.

Williams mostly focuses on reasons for belief. He has less to say about the source of our reasons for desiring certain things or preferring one thing to another. For instance, in the above example, the perverse interpretation gratuitously assigns Prelinguistic Sally an unreasonable desire for severe pain. What makes this desire unreasonable? In general, what makes certain final desires reasonable and others unreasonable?

Here again Williams might appeal, at least in part, to Prelinguistic Sally's conscious experiences. Sally has a range of experiences with 'valence': pleasures, pains, gustatory experiences, and feelings. The phenomenal contents of these experiences (attributing qualities to bodily regions) are certainly internally-determined, rather than being determined by teleological relations to environment states. Just as it may be a basic fact that visual experiences give Sally reasons to believe various things, maybe it is just a basic fact that affective experiences with certain phenomenal contents give her a reason to desire certain things to varying degrees. If so, then the 'best interpretation' will be one that tends to assign her desires that are 'reasonable' given her affective experiences.

**3. Subject-Based or State-Based Radical Interpretation Theory?**

Williams notes (p. 33ff) that there are two ways of further developing the radical interpretation theory: *subject-based* and *state-based*.

Let us start with the subject-based version. And let us continue to work with Prelinguistic Sally.

> **Subject-based radical interpretation theory.** An interpretation assigns beliefs and desires in the first instance *directly to Prelinguistic-Sally-at-a-time (a person-slice)*, based on her experiences and consequent dispositions to act *at that time*.

Williams gives an *a priori* argument that subject-based radical interpretation theory *fails for any possible subject S* (p. 33ff). Take any possible subject *S* who has beliefs and desires. Now consider a *Blockhead-duplicate* of *S* (named after Ned Block who first described the example). Blockhead is a mere robot that has the same physical input-output dispositions as *S*, but not the same internal organization. In particular, Blockhead works by a giant look-up table. If the subject-based version of Williams' radical interpretation theory is right for *S*, then apparently this Blockhead duplicate of *S* will have the same beliefs and desires as *S* (aside from differences in wide content). But, intuitively, Blockhead is a mere unintelligent machine lacking beliefs and desires.

In response to Blockhead, Williams proposes:

> **State-based radical interpretation theory**. For any possible subject *S*, *S* has a core belief that *p* iff *S* is in some repeatable internal (e. g. neural) state *N* and, given *N's overall* functional role in *S* in the past and present with respect to experiences and behavior, the most rationalizing overall assignment assigns to *N* the belief that *p*.

This avoids the Blockhead counterexample. It implies that, as a matter of *metaphysical necessity*, if *S* has beliefs, *S* decomposable into separate states that are assigned those beliefs. No Blockhead satisfies this condition.

Note that Williams state-based radical interpretation theory is very strong in the following sense. He distinguishes between his foundational theory and contingent details of realization (xvii-xxi). The foundational theory consists in metaphysically necessary biconditionals. He suggests that the state-based formulation is meant to have this status (p. 34-35). It is arrived at *a priori* and is meant to rule out any actual *or possible* Blockhead believer.

There is an issue for the state-based account that Williams mentions but does not say much about. He writes:

> I'll be assuming that the attitude-types [of state-tokens] (e.g. flat-out-belief, supposition, degree of belief, degree of desire) are grounded prior to and independently of the determination of the contents that they are paired with. (p. 40; see also p. 45)

This suggests that Williams assumes a Fodor-style, *two-part story* for believing that *p:* to believe that *p* is to be in an inner state that (i) is a belief (because it plays the 'belief-role') and that (ii) is assigned content *p* (Fodor 1978).

In fact, Williams' view is Fodorian in another way: he assumes that, in actual humans at least, the relevant internal states have syntactic structure: they are sentences in a Fodorian inner language of thought that we cannot introspectively access (p. 11, 40, 50, 156). But whereas Fodor assigned contents to the inner sentences based on 'asymmetric dependence' relations to external states, Williams' radical interpretation theory assigns contents to them based on 'rationality maximization'.

Is Williams correct that a strong, metaphysically necessary state-based version of radical interpretation theory is to be preferred to the more neutral subject-based version, because it rules out possible Blockhead believers? One question here is whether it is not too strong. It may also rule out believers we take to be possible. For instance, imagine *Connectionist Sally*. She is a possible agent who has conscious experiences of the world and acts on that world, but she works by a holistic, connectionist neural network that isn't an

implementation of state-based architecture (Stich 1996, chapter 2). Or imagine *Non-physical Sally* (Fodor 1978, pp. 520-1). She has conscious experiences, and they modulate her decisions, but she lacks a rich system of underlying mediating (physical or non-physical) states. Intuitively, unlike Blockhead, Connectionist Sally and Non-Physical Sally might have many beliefs and desires. But, because they are not decomposable into distinct corresponding states that are assigned those beliefs and desires, Williams' strong, metaphysically necessary version of the state-based radical interpretation theory rules out such believers no less than Blockhead believers.

There is another potential worry about Williams' state-based version of radical interpretation theory. It allows for counterintuitive 'absent role' cases of belief. For instance, let *N* be the internal physical state that, *throughout Prelinguistic Sally's life*, was typically caused by the experience of a lion and typically caused avoidance behavior (etc.), so that the best interpretation pairs it with *the belief that a lion is present*. However, late in her life, something odd happens on a single occasion: *N* is momentarily caused by a pebble on the ground and causes her to reach for the pebble. It is not caused by an experience (or hallucination or imagination) as of a lion being present. And it is not apt to cause any behavior appropriate to a lion being present (including linguistic behavior, since Sally lacks an outer language). Nevertheless, a state-based radical interpretation theory apparently implies that, by virtue of being in internal state *N* on this occasion, she momentarily has an entirely secret and irrational *belief that a lion is present* – even though for all the world she merely believes that a pebble is present. For the *overall* most rationalizing assignment of beliefs to Sally's internal states will assign to the repeatable state *N* the belief that a lion is present, even if *on this occasion* it implies that Sally has a deeply irrational belief. But, intuitively, this is the wrong verdict about this case. In this case, Sally just doesn't secretly acquire and then lose a secret belief that a lion is present, because all the while she has absolutely no experiences or dispositions congruent with that belief. Mad *pain* may be possible (Lewis 1980), but this extreme case of mad belief – this 'secret scrambling' – is impossible. By contrast, a subject-based radical interpretation theory avoids attributing the momentary lion-belief to Sally. For this version assigns beliefs and desires directly to Prelinguistic-Sally-at-a-time, based on her experiences and consequent dispositions *at that time*, which are not congruent with the lion-assignment. This supports a subject-based version of radical interpretation theory for Prelinguistic Sally's beliefs and desires.

In response to these worries, Williams might acknowledge that the state-based requirement should not, after all, be built into his foundational, metaphysically necessary story for belief and desire. This would allow Connectionist Sally and Non-physical Sally to have beliefs and desires. But

then he would need another way of ruling out the possibility of Blockhead believers. One idea is that he might add a 'causality condition' to the theory (for this idea, and for a general development and defense of a subject-based rather than a state-based view, see Braddon-Mitchell and Jackson 2007, p. 119ff, 197; but see O'Rourke 2018, chapter 4 for a problem with the 'causality condition'). Another idea would be to move to a more 'consciousness-first' radical interpretation theory of belief and desire as discussed above. This might rule out Blockhead believers assuming that Blockheads are mere insentient automata lacking conscious experiences.

**4. Williams' Mind-First Approach to Linguistic Representation**

So far, we have imagined that Sally is one of our pre-linguistic ancestors. Now let us imagine something fanciful: Sally manages on her own to invent an outer language, which gradually becomes more and more sophisticated. She develops a base-ten number system, logical words, and increasingly abstract words of all kinds.

What determines the representational properties of Sally's language? Williams develops a Lewisian view with two parts:

> **Belief is prior to linguistic content**. On Williams' form of the radical interpretation theory, the account of what grounds Sally believing that *p* never involves her accepting an outer sentence that means that *p*. In fact, it never involves facts about the *meaning* of outer language at all. In this sense, 'mental content is metaphysically prior to linguistic content' (pp. 146-147). Therefore, it can be used to *explain* linguistic content, as follows.
>
> **Language inherits content from belief.** Sally enters into conventions associating a finite stock sentences (the sentences she actually utters) with the contents of her beliefs. These 'data points' help determine the correct compositional meaning theory for Sally's whole language (including never-uttered sentences): it is the *simplest*, *most elegant* theory that fits with the data (p. 149ff).

Williams develops an interesting form of social externalism that is in line with his thesis that mental content is metaphysically prior to linguistic content (p. 140ff). But we can pretend that Sally is on her own, so that his social externalism may be ignored in this case.

Now one issue for Williams' approach to linguistic representation is: what does 'the simplest' compositional meaning theory mean? Lewis tried to define simplicity in terms of length-of-definition in 'ontologese': a language only containing terms for fundamental (physical) properties and relations. In a departure from

Lewis, Williams proposes a subject-sensitive account on which the simplicity of a compositional meaning theory for the outer language of subject *x* is defined in terms of 'minimal descriptive length' in the basic terms of *x's inner language of thought*. So, as he says (p. 156), he 'presupposes [an inner language of thought] in the very formulation of the account of linguistic content'. But a wrinkle Williams doesn't address is that, since the language of thought hypothesis is contingent, there might be an individual who lacks an inner language of thought but who has an outer language for which there is a true compositional meaning theory. What is the alternative account of simplicity in this case?

There is another, deeper issue for Williams' general view that mental content is explanatorily prior to linguistic content. Although we have now imagined Sally to have developed a sophisticated outer language, let us return to her more primitive, prelinguistic stage. Prelinguistic Sally had many beliefs and desires about her environment. But there were limits. Given the range of experiences and behaviors they are capable of, humans who have never had any outer language simply cannot believe certain things: for instance, that the laws of quantum mechanics are so-and-so, or that the number of grains of sand in a certain heap is precisely 6,483,074. For instance, no possible course of Prelinguistic Sally's human behaviors and human sensory-perceptual experiences could make it the case that she has such beliefs. In order to have such beliefs, Sally needs access to an outer language that can express these things. This idea – call it *prelinguistic limits* – is a common one in philosophy and cognitive science.

Prelinguistic limits creates a *prima facie* explanatory challenge for Williams' general thesis that mental content is prior to linguistic content. By this thesis, the metaphysical ground of a normal human *believing* that *p* will never involve an outer sentence *meaning* that *p*. This suggests that, for nearly any belief that *p*, a normal human could in principle have the belief that *p without* having access to any outer sentence or representation that means that *p*. (Of course, beliefs *about* outer language would be an exception.) So one challenge for Williams is this: given his priority thesis, why should *any* beliefs require an outer language? He also faces a contrastive explanatory challenge: why do *some* beliefs require an outer language while *others* do not?

In response, Williams might insist that he can explain prelinguistic limits in a way consistent with his thesis that mental content is prior to linguistic content. Perhaps the explanation is that in some cases *linguistic behavior* is an essential part of the total set of behaviors the rationalization of which

grounds attributing a certain belief. This should not be understood as the claim that an essential part of the ground of believing certain things is accepting sentences that *mean* those things. For, in that case, the claim would be inconsistent with Williams' view that linguistic content is never prior to mental content. Rather, the claim is that an essential part of the ground of believing certain things is 'linguistic behavior' *understood non-semantically:* for instance, making certain types of noises (thanks to Williams here).

But this explanation of prelinguistic limits raises some questions. Given Williams' radical interpretation theory, *why* should it be that linguistic behavior, understood as merely making certain types of noises, is an essential part of the ground of *any* beliefs? And why should it be an essential part of the ground of having *some beliefs but not others?* How do these assertions flow from his general radical interpretation theory? Second, the explanation only asserts a necessary connection between Sally having certain sophisticated beliefs and her merely *making certain types of noises.* But, intuitively, what needs explaining is instead a connection between certain beliefs and linguistic *content:* a human's believing certain sophisticated contents requires their having access to some outer representations or other that *express* those contents.

There is another response available to Williams. In order to explain prelinguistic limits, he might jettison his general claim that mental content is always prior to linguistic content and move to a mixed view (e. g. Speaks 2010, p. 234). For example, one idea is that for a subject to believe that *p* is *either* for the subject to be assigned this belief by a simple subject-based radical interpretation theory (in terms of experience and action) *or* for the subject to 'accept' or 'believe*' an outer language sentence that means that *p* (where acceptance is explained in functional terms). Thus some beliefs are language-mediated while others are not. This mixed view might elegantly explain prelinguistic limits. For example, if Prelinguistic Sally's sensory-perceptual experiences can only have quite thin contents and her reasons are therefore quite limited, the radical interpretation ('rationality maximization') theory is only able to assign to her determinate beliefs and desires within certain limits. The only way for Sally to believe more sophisticated contents outside these limits (about quantum mechanics, large exact numbers) is by satisfying the second disjunct, requiring that she have access to an outer language capable of expressing those contents

Even though this mixed view would require that Williams reject his thesis mental content is always prior to linguistic content, it might retain the main elements of his view. As noted above, Williams assumes a 'language of thought' approach: a person has a (core) belief that *p* iff the person believes* an *inner* sentence that means that *p*, where believing* is explained in functional terms. To explain what it is for an inner sentence to mean that *p*, Williams appeals to a kind of rationality-maximization theory of content-

determination (he focuses on logical connectives and moral terms in the inner language of thought). If he were to move to the suggested mixed view, then Williams would retain the language of thought approach in the case of sophisticated beliefs, but with one tweak: the relevant language is the outer language (e. g. English) rather than a hidden inner language. He might also retain his rationality-maximization theory of content-determination, but simply apply it directly to outer language rather than to an inner language (if such there be).

**5. Conclusion**

As should be clear by now, Williams' book is an important contribution. It is the kind of 'big picture' book that makes one excited about philosophy. But it also supplies the details. In fact, it is filled with many new and ingenious ideas that I haven't discussed (e. g. a derivation of reference magnetism, an explanation of the referential stability of moral terms, and much else). It is essential reading for those interested in the foundations of the mind's capacity to represent the word.•

ADAM PAUTZ
*Brown University, USA*
adam_pautz@brown.edu

**References**

Braddon-Mitchell, D. and F. Jackson. 2007. *Philosophy of Mind and Cognition.* Oxford: Blackwell.

Fodor, J. 1978, Propositional Attitudes. *The Monist* 61: 501-523.

Lewis, D. 1980. Mad Pain and Martian Pain. In N. Block (ed.) *Readings in the Philosophy of Psychology.* Harvard University Press. pp. 216-222.

O'Rourke, J. 2018. *Property Dualism as a Solution to the Mind-Body Problem.* Princeton Dissertation.

Pryor, J. 2000. The Skeptic and the Dogmatist. *Noûs* 34: 517-549.

Speaks, J. 2010. Explaining the Disquotational Principle. *Canadian Journal of Philosophy* 40: 211-238.

Stich, S. 1996. *Deconstructing the Mind*. Oxford: Oxford University Press.

• I am grateful to Robbie Williams for many helpful discussions and to Brian Cutter for comments on an early version of this review.