BOOK SYMPOSIUM

Varieties of Interpretationism about Belief and Desire

ADAM PAUTZ

In his superb book, *The Metaphysics of Representation*, Williams sketches biconditional reductive definitions of representational states in non-representational terms (xvii).[1] The central idea is an extremely innovative variety of *interpretationism* about belief and desire. Williams is inspired by David Lewis but departs significantly from him.

I am sympathetic to interpretationism for some basic beliefs and desires. However, I will raise *three worries* for Williams's version (§2–4). Then, I will suggest a modified version (§5). I will conclude with a general question (§6).

1. *Williams's multistage interpretationism*

To illustrate Williams's account, imagine that we have travelled back in time and are radically interpreting *Sally*, one of our prelinguistic hominid ancestors.

Donald Davidson's interpretationism started from an agent's disposition to 'hold-true' certain sentences of her public language, so that prelinguistic Sally is not interpretable as having beliefs and desires at all. Against this, Williams (xi) holds that she does have beliefs and desires. But how do we pin down their contents?

The first stage of Williams's theory (167ff) concerns *source intentionality:* prelinguistic Sally's perceptions and decisions. For example, suppose that there is a ripe tomato in front of her and she picks it from the vine. Her perception has the content *there is red and round thing there*. Her decision has the content *I will move my hand in direction d*.

Some hold that source intentionality is a matter of internally determined 'experiential intentionality'. Williams considers this idea but rejects it (13, 168). Instead, Williams explains source intentionality using Karen Neander's externalist-teleological theory of perceptual content. Sally undergoes an inner physical state *N* (a functionally defined 'perception') that has the biological function of being caused by a round thing with a red-reflectance. This gives her a reason to believe that such a thing is there. Sally also undergoes inner physical state *A* that has the function of causing her to move her hand in direction *d*.[2]

1 All references are to Williams 2020 unless otherwise noted.

2 Williams holds that source intentionality is quite 'low-level' or 'thin' (183, 178). Williams might also include memories (perhaps identified with spontaneous judgements) in source intentionality (181, n. 16).

Analysis Vol 00 | Number 0 | 2021 | pp. 1–13 doi: 10.1093/analys/anab044

In the second stage, Williams goes from source intentionality to Sally's beliefs and desires. In the case of belief, the idea is:

> **Williams's interpretationism.** For any possible subject *A*, *A* has a (core) belief that *p* iff *A* is in a distinct repeatable inner state *s* such that (i) *s* plays the belief-role and (ii) the most rationalizing interpretation (given source intentionality) assigns to *s* the content *that p* (33).

In the case of actual humans, Williams's account requires at various places that the relevant internal states are sentences in an *inner* language of thought (LOT) that she cannot experience (11, 40, 50, 156). So while prelinguistic Sally lacks an *outer* language that she can experience, she has a hidden inner language.

This resembles Jerry Fodor's two-part story: a person believes that *p* just in case (i) inner sentence *s* is in their 'belief-box' and (ii) *s* means that *p* (33, 40).[3] But whereas Fodor explains (ii) in terms of 'asymmetric dependence', Williams explains (ii) using rationality maximization.[4]

What is the 'rationality maximizing' assignment? It is the assignment that overall maximizes Sally's structural rationality (e.g. means-end coherence) and substantive rationality (e.g. reason-responsiveness), given source intentionality (16ff).

For example, let's grant that prelinguistic Sally has a hidden LOT. When Sally views the tomato and picks it from the vine, there are two inner sentences s_1 and s_2 that mediate between her perception and her action in a way distinctive of a belief and a desire. One perverse interpretation rationalizes her action by assigning the (clearly false) content *a chunk of mud is there* to s_1 and the content *I will eat a chunk of mud* to s_2. But the interpretation that also maximizes Sally's substantive rationality will instead assign the contents *a tomato is there* and *I will eat a tomato*. So Sally believes that a tomato is there and she desires that she will eat it.

The third and final stage concerns the content of *outer* language (119ff). To illustrate, suppose we determined the contents of prelinguistic Sally's beliefs and desires. Presumably, they are limited. Then something new happens. Sally and her group invent an outer language. The language becomes increasingly sophisticated. At the same time, she comes to have increasingly sophisticated beliefs and desires.

What determines the representational properties of Sally's outer language? This raises one of philosophy's chicken-or-egg problems: in the order of explanation, what comes first, thought or outer language? Evidently, Sally's simple beliefs are prior to outer language. Williams (146–47) endorses a stronger claim: *all* belief content is metaphysically prior to the content of outer

3 Williams does not discuss part (i) of the two-part theory. He does not offer a general functional account of what it is for an inner sentence to be in Sally's belief-box or desire-box.

4 Williams focuses on certain sub-sentential terms of LOT (e.g., logical constants, moral terms). He has less to say about what determines LOT's grammar and hence the full contents of all of LOT's sentences.

language (even if they develop together). Call this a *general mind-first approach* to belief as opposed to an *outer-language-first* approach. Then, following David Lewis, Williams derives the content of outer language from the content of belief, *via* conventions (119ff).

Why think that constraints of rationality are *constitutive* of content? Williams offers no master argument. If his theory delivers correct verdicts, it may be along the right lines.

2. *First worry: neglecting consciousness*

Prelinguistic Sally views a tomato. Following Karen Neander, Williams holds that Sally's perceptually representing *that a round thing is there* is a *wide* affair that is not fixed by her total intrinsic state. Furthermore, by having a perception with this wide content, she has a reason to believe that a round thing is there. So the most rationalizing interpretations will tend to assign to her the belief that a round thing is there. In short, his account of how perception helps fix belief can be put like this:

wide content → reason → belief

But the following argument suggests that Neander-style wide content cannot be the whole story:

(1) When she views the tomato, Sally's conscious experience has a content *there is a reddish and round thing there* that is inseparable from its phenomenology.
(2) Phenomenology is narrow.
(3) Therefore, in addition to having a wide content determined by causal-teleological relations to the environment, Sally's experience has a narrow *experiential* content that is inseparable from its phenomenology.

Premiss 1 is based on reflection on experience and is generally accepted. Phenomenal externalists reject Premiss 2. But research in psychophysics and neuroscience provides an overwhelming case for it. So perceptual source intentionality includes narrow experiential intentionality (Pautz 2013, 2021a).

Williams's reductive externalist account borrowed from Neander does not apply to such internally determined experiential source intentionality. So if Williams wishes to uphold biconditional reductive definitions of all representational facts (xvii), he needs an alternative internalist reductive account here: subject *S experientially represents* narrow content *p* iff ___. This has proven difficult to provide, because standard models for reducing representational relations are externalist.[5]

5 Williams's interpretationism invokes normativity. He allows that normative facts may be grounded in physical-natural facts even though biconditional reductions are unavailable (13). So why not a parallel view of experiential representation? Perhaps the fact that Sally

The argument puts pressure on Williams's view about the source of perceptual reasons. For, intuitively, Sally has a reason to believe that a *round* thing is there simply by virtue of *experientially representing* that a round thing is there, where this is matter of how the world phenomenally appears to her. The argument shows that this is a narrow affair. By contrast, on Williams's view, Sally has a reason to believe that a *round* thing is there by virtue of *Neander-representing* that a round thing is there, where this is wide affair involving her being in an internal physical state that is normally caused by a round thing in the external world. Since phenomenology is narrow, this is totally independent of how the world phenomenally appears to her. So Williams's view becomes less plausible. At least, he neglects a plausible 'narrow' source of perceptual reasons.

So here is another interpretationist model of how perception fixes belief:

narrow experiential content → reason → belief

Sally's internally determined conscious experiences (pleasures, pains, gustatory experiences and feelings) may also play a role in providing Sally with reasons for *desiring* certain things or *preferring* one thing to another (Pautz 2021a: 288–89, 2021b). The 'best interpretation' will then assign to her desires that are 'reasonable' given her affective experiences.

At the end of §1, I asked, 'Why accept Williams' view that substantive rationality plays a role in grounding belief and desire?' Let me mention one explanatory virtue. Intuitively, if it experientially appears to you that a red and round thing is *right there*, you are apt to believe that such a thing is there – that content becomes a belief magnet. It is then very hard not to believe this content. Likewise, if you have horrible pain, you are apt to desire that it stop. And, intuitively, such connections are not merely contingent but metaphysically necessary. But why? This can be explained by the reason-grounding role of consciousness together with Williams's reason-responsive theory of beliefs and desires, as follows: (i) it is in the essence of your conscious experiences to provide reasons for beliefs and desires and (ii) it is in the essence of your beliefs and desires that they tend to be congruent with your reasons (for in the absence of countervailing behavioural dispositions the 'most rationalizing' interpretations will assign beliefs and desires congruent with your reasons).

3. *Second worry: do beliefs and desires depend on hidden facts?*

Theories of beliefs and desire fall into two vague categories. On one side are theories that mostly ground beliefs and desires in surface-level facts. By *surface-level facts*, I mean a rough miscellany that includes not only facts accessible from the outside but also experiences, imagery, conscious doings,

experientially represents so-and-so is *grounded in* the fact that she is in a certain narrow brain state without being *reducible to* that fact. All representational facts are grounded in (but not reducible to) non-representational facts.

acceptances of outer sentences inner speech, and causal-inferential relations between these things. Gilbert Ryle's view was a paradigmatic example. Recent phenomenal intentionality theories are also in this camp (Chalmers 2012: 463–67, Mendelovici 2018, Pautz 2021a). On the other side are theories that significantly appeal to 'hidden facts'. Fodor's hidden LOT view is an example.

Williams holds that certain hidden facts are crucial to the grounding story. This makes it open to two problems.

3.1 First problem

Williams's *state-based* form of interpretationism (§1) implies the *inner isomorphism constraint:* as a matter of *metaphysical necessity*, an agent has beliefs only if there is a one-one mapping between those beliefs and a system of inner states (representation-vehicles) causally mediating between experiences and behaviours (33–34).[6] Likewise for desires. These mediating states will be subpersonal and hidden. For instance, when an animal experiences the world and acts on it, the system of mediating inner states realizing its beliefs and desires are not experienced. Williams gives an *a priori* argument for the constraint: any possible *Blockhead* (an insentient robot that outwardly looks like a human but that works by a giant look-up table) lacks beliefs and desires only because it violates the constraint (33–35). Therefore, he rejects the Lewis–Stalnaker view that a single time-slice of an agent might be assigned many beliefs and desires (33).

I think that Williams's inner isomorphism constraint is too strong. Here are two illustrations.

Imagine that you are a primitive pre-human of our actual past – say *homo habilis* – with many basic beliefs and desires. In fact, you satisfy the inner isomorphism constraint: you have a distinct brain state for each one. But now imagine a twin in another possible world with the same experiences and behavioural dispositions as you. The only difference is that, in your twin, they are mediated by a connectionist network. We stipulate that the network is not decomposable into distinct states that are 'isomorphic' to your own mediating brain states here in the actual world, even at an abstract upper-level. So Williams's constraint implies that your *homo habilis* twin lacks beliefs and desires for the same reason it implies that Blockhead does: he fails to satisfy the inner isomorphism constraint. But that is implausible. For instance, when your twin has a vivid experience of a tomato and reaches for it, your twin surely believes that a red and round thing is there. Here is an intuition pump: if you could cycle between having the first architecture and the second (say, once every hour), you would not even notice a difference, and

6 I assume that Williams's constraint is that each distinct belief be realized by a *distinct* underlying representation-vehicle, for three reasons. First, Williams says that having the same beliefs requires 'isomorphic' inner states (34). Second, he requires that distinct beliefs can play distinct causal roles (34–35). Third, without 'distinctness', the constraint becomes compatible with the Lewis–Stalnaker single-state/time-slice view that he uses Blockhead to argue against. (Incidentally, Williams would presumably restrict the constraint to *core* beliefs.)

your dispositions in all circumstances would remain the same. It is implausible that you would unnoticeably cycle between having beliefs and desires and having none at all. Your beliefs and desires would remain the same.

Here is another example. Consider you as you are now. Now imagine a distant counterfactual situation in which you have a twin. All the surface-level facts about your twin described without using 'belief' and 'desire' are the same: same inner experiences, same dispositions to behave, same use of outer language in inner and outer speech, same *causal relations* between these things, same interactions with things and people in the outside world, and so on. If you could cycle back and forth between the actual and hypothetical situations, you would not notice a difference. But there is a difference in the underlying metaphysics. In the hypothetical situation, interactionist substance dualism is true of your twin. Indeed, he lacks a brain (even a Blockhead-style look-up table). Instead, inputs to your twin's sensory organs directly cause his experiences and his experiences directly influence his behavioural outputs. (This need not involve agent causation.) So your twin's experiential mental life is the same as yours but it is ungrounded – it is not even grounded in underlying non-physical states.[7] Williams's inner isomorphism constraint again implies that your twin cannot have the same beliefs and desires as you. But this again is implausible. For instance, when your twin experiences a tomato and so reaches for it, he surely also believes that a red and round thing is there, even if this belief is not realized by an inner representation-vehicle (not even a non-physical soul-state) mediating between experience and action. And, intuitively, your twin can believe *that either snow is white or purple pigs can fly* by virtue of understanding and accepting (in a functional sense) the *outer* sentence 'either snow is white or purple pigs can fly', even if there is no underlying inner state (not even a non-physical soul-state).[8] If you discovered that *you* are such a brainless non-physical subject directly interacting with your body, you should not conclude that you lack such beliefs!

3.2 Second problem

So, Williams's view implies that your having beliefs and desires *at all* is implausibly hostage to hidden facts. Next we will see that it also implies that *what* you believe depends on hidden facts.

7 Even if physicalism is actually true (so that our experiences are actually grounded in something more basic), it is not necessarily true. The scenario certainly cannot be excluded *a priori*.

8 For Williams causal-inferential basing relations are important to interpretation (35, 36ff). But your Cartesian twin accepts some outer sentences (e.g. 'either snow is white or purple pigs can fly') on the basis of accepting others (e.g. 'snow is white'). It is just that, in him, these causal transitions are not grounded in anything more basic. So Williams might assign contents directly his *outer* language, and ground some of his beliefs in his accepting outer-language sentences. This leads to 'modified interpretationism' (§5).

Here is one example.[9] Recall that Williams holds that, as a contingent matter, the relevant inner states of actual humans are sentences of an inner LOT. Suppose that this is right. And suppose that at 10 am one day, for only 10 seconds, your hidden inner LOT terms 'red' and 'green' were interchanged throughout your cognitive system.[10] But you had no idea, because there was no change in your surface-level experiences and dispositions: you had the same experiences of grass and tomatoes, the same dispositions to accept (in inner and outer speech) outer language sentences like 'the grass looks green', 'by *green* I mean THAT quality', 'the tomato looks red' and so on. In fact, at the time, you yourself were disposed to insist 'I didn't change my beliefs about what colour qualities things appear to have!' Given all this, it is quite clear that there was no secret change in your beliefs about what colour qualities things appear to have – your beliefs stayed constant while being differently realized.[11] By contrast, Williams's specific inner sentence variety of interpretationism implies that, for 10 seconds, you secretly had radically different (and radically mistaken) beliefs about what colour qualities things appear to have. For he often says (43) that the correct interpretation of LOT terms is the one that is most rationalizing *overall*, which might imply extreme irrationality in an isolated case.

There are, then, problems for Williams's view that hidden facts are a crucial part of the grounds of beliefs and desires. Williams might consider an alternative hypothesis: the minimal grounding base (and scrutability base) for your beliefs and desires need not include certain hidden facts about realization involving inner isomorphism and LOT. Radical interpretation does not need them. To undermine this hypothesis, one must show that discoveries about those specific hidden facts could make a difference to your beliefs and desires while holding all else fixed, including all surface-level facts. But the above cases show that this is implausible. That supports the surface-level hypothesis. If it is right, then, while you likely have a hidden inner LOT, it is not part of the minimal grounding base. Instead, outer language may play an important grounding role.

9 This example was suggested to me by David Chalmers. For another example, see Pautz 2021b.

10 This should be possible. After all, when the apparent colors of things switch, this causes your inner LOT color terms to switch in kind. Why could not your LOT color terms switch without this cause?

11 If in your sleep you become mentally ill or your brain is tampered with, you might wake up with radically different and irrational (dispositional) beliefs, without noticing at first. But here your *dispositions* to say and do things will also change. By contrast, in the red-green interchange case, your experiences and surface-level dispositions *under all circumstances* are stipulated to remain the same. So your beliefs clearly do not radically and secretly change.

4. *Third worry: against a general mind-first approach to belief*

In §1, I noted that on Williams's variety of interpretationism 'mental content is metaphysically prior to linguistic content' (146–47).

Williams does not give an argument for this. He does note (xi) that individuals lacking outer language can have simple beliefs. But this does not show that *all* belief content is metaphysically prior to the content of outer language. There is another option Williams does not consider: a pluralistic variety of interpretationism that is mind-first for simple beliefs and outer-language-first for sophisticated beliefs (§5).

In my view, there is no convincing argument for Williams's contrary mind-first view that all belief content is metaphysically independent of linguistic content. On the other hand, there is a strong argument against it. Intuitively, there are sophisticated contents a human cannot believe if she lacks access to an outer language expressing those contents.

To begin with, consider some things humans (ordinary people, physicists, philosophers) can explicitly believe with outer language:

(a) There are at least 1,920,762,973 people on Earth.
(b) Either so-and-so quantum mechanical laws are true or else there exists an alien riding on a green zebra while solving differential equations.
(c) Some people believe that they believe that within the inner sphere of possibility everything supervenes on the physical.

Could any normal *prelinguistic* human believe (a)–(c)? I say no. (By a 'normal' human here and in what follows, I mean a human with *normal human source intentionality*: the normal range of human experiences with relatively thin phenomenal contents.) For instance, imagine Sally as one of our normal prelinguistic human ancestors. Prelinguistic Sally's group lacks any *outer* language – meaning any language that they can experience. So prelinguistic Sally is also incapable of inner speech. Now just try to describe possible cases in which prelinguistic Sally plausibly believes (a)–(c). You cannot do it. Without an outer language, whatever she experiences or does, it will be insufficient to ground her believing (a)–(c) rather than some other contents.[12] Most accept that there are limits on what nonlinguistic *animals* can believe. Why? Because we cannot describe a possible case in which they believe such things. For the same reason, we should accept the same for prelinguistic *humans*.

Now here is why this creates an argument against Williams's general mind-first view that all belief content metaphysically independent of linguistic content:

12 Perhaps a hypothetical super-experiencer could have an experience in which it phenomenally seems to her that there are exactly 1,920,762,973 grains of sand in a pile (or whatever), and so could believe ('subitize') this content by simply endorsing the content of her experience, without needing outer language. But this is irrelevant, since here and in what follows I am restricting the claim to a 'normal' prelinguistic human (i.e., one with normal thin source intentionality).

(1) Williams's general mind-first theory of belief content cannot plausibly explain why a normal prelinguistic human could not believe (a)–(c); indeed, it arguably implies that such a human *could* believe (a)–(c).
(2) An alternative pluralist theory (§5) can more plausibly explain why this cannot happen.
(3) This favours such a pluralist theory over Williams's general mind first approach.

The case for Premiss 1 is simple. Williams's mind-first view holds that the metaphysical grounds of a normal human believing (a), (b) or (c) in no way involves the semantics of outer language; therefore, it does not involve their having access to an outer language that can *express* (a), (b) or (c).[13] If so, then, whatever those grounds are, they could in principle be present in a normal human that lacks access to an outer language that can *express* (a), (b) or (c). Therefore, the mind-first approach implies that a normal prelinguistic human (e.g. prelinguistic Sally) could in principle believe (a), (b) and (c).[14]

Williams might try to resist Premiss 1. He might agree that it is humanly impossible that a normal prelinguistic human should believe (a), (b) or (c). But he might try to explain this is in a way consistent with his mind-first approach. But elsewhere (2021b) I present grounds for scepticism.[15]

13 Williams defends a form of social externalism but, as he notes (137, 145), it is compatible with his 'mind-first' claim. Indeed, on Williams's view (142), a human alone on an island who invents a 'private language' could come to believe (a), (b) and (c) without outer language being involved in the grounding story at all.

14 Of course, prelinguistic Sally could not believe (a) by way of our linguistic mode of presentation '1,920,762,973'. But, by the argument in the text, Williams's view does imply that prelinguistic Sally could in principle have beliefs with contents (a), (b) and (c) where they are understood truth-conditionally or in terms of structured Russellian propositions. (Thanks to Williams here.)

15 Williams accepts a mind-first account on which an *inner LOT* is part of the grounding story. Given this view, it is especially difficult for him to explain why prelinguistic Sally could not believe (a), (b) or (c). For, in principle, she could have normal source intentionality and a very rich hidden inner language, whose terms play certain complex functional roles. Why could not that be enough to ground her believing (a), (b) and (c)? Why would she *also* need an outer language? For example, in discussion, Williams suggested to me that an outer language may be needed to be 'sensitive to certain distinctions'. But, in principle, why could not a rich *inner* language be so connected to the outside world as to enable prelinguistic Sally to be sensitive to those distinctions? If Williams instead accepted the more 'surface-level' interpretationist analysis to be discussed in §5, he could explain why outer rather than inner language is required. On that view, the concept of belief is only sensitive to the surface-level facts about an agent (§3). So even if prelinguistic Sally has a rich subpersonal inner language, it is irrelevant to the grounding story. In her case, the relevant surface-level facts only include her experiences, imagery, behavior and so on. And those surface-level facts will always be insufficient to determine that she believes (a), (b) or (c). On a surface-level analysis, only *outer* language will do the trick.

Next consider Premiss 2. I will now suggest a modified pluralist interpretationism that can plausibly explain the limits on prelinguistic thought. It also avoids the worries I raised in previous sections.

5. *Modified interpretationism?*

The modified form of interpretationism I will suggest retains Williams's basic structure: source intentionality and then interpretationism for belief and desire. Very roughly:

> **Modified interpretationism.** *A* believe that *p* iff *either A* is directly assigned the belief that *p* by the most rationalizing interpretations given source intentionality *or A* is disposed to 'accept' (in outer or inner speech) an *outer* language sentence that is correctly interpreted as meaning that *p*.

To illustrate, imagine that Sally belongs to a prelinguistic tribe. By virtue of her conscious experiences and dispositions to act, she counts as believing a limited range of simple contents about her environment in the first way. Then her group invents an outer language that becomes increasingly sophisticated, referring to objects and properties far outside their perceptual circle. They retain their simple, mind-first beliefs. But, intuitively, they now have a new way of believing more sophisticated contents: by understanding and accepting outer language sentences correctly interpreted as expressing those contents.

As Dennett (1987: 233) says, 'there are really two sorts of phenomena being alluded to by folk-psychological talk about beliefs: the verbally infected states of language-users and the deeper states of what one might call animal belief'. Put differently, there is a single phenomenon with a plurality of grounds.

Let me explain how modified interpretationism might allow Williams to avoid the three worries I raised.

First, modified interpretationism might include Sally's richly intentional, internally determined conscious experiences within 'source intentionality' (§2). This source intentionality helps fix her language-independent beliefs (first disjunct) and also helps to fix the content of her outer language (second disjunct).[16]

Second, modified interpretationism only appeals to broadly surface-level facts. Instead of assigning contents to her inner LOT, it assigns contents directly to herself or to outer sentences she accepts, based on surface-level facts (her experiences and dispositions, her and her community's use of outer language). This would allow Williams to accommodate the plausible hypothesis (§3) that broadly surface-level facts are a minimal grounding base for her beliefs and desires.

16 I have developed this multistage interpretationism (roughly, experiential source intentionality and then interpretationism for the rest) in Pautz 2013 and 2021b. Source intentionality need not *only* include narrow experiential content; it can also include wide facts about what external objects and kinds Sally counts as perceiving and acting on.

Third, modified interpretationism would allow Williams to neatly explain another intuitive datum (§4): no matter what she does, a prelinguistic human with the normal range of experiences cannot believe extremely sophisticated contents such as (a), (b) or (c).

For instance, consider prelinguistic Sally back when she belonged to a primitive tribe. Suppose she has a magical subpersonal mechanism that can detect only one condition: exactly 1,920,762,973 grains of sand before her. When this happens, it causes her to decide to scratch her head, apparently out of the blue. But she has normal human experiences. So it does not then experientially appear to prelinguistic Sally *that exactly 1,920,762,973 grains of sand are there*. It just experientially appears to her *that a whole lot of grains of sand are there*.

To rationalize her head-scratching, should we attribute to prelinguistic Sally the belief that exactly 1,920,762,973 grains of sand are before her, and the desire to scratch her head when this is so? Intuitively not. She herself has *no idea* exactly how many grains are sand are there (even if her *subpersonal* mechanism detects this precise number) – just as you would have no idea.

Modified interpretationism can explain this. The correct assignment of beliefs and desires to prelinguistic Sally must be congruent not just with her dispositions to act but also with the reasons provided by her *conscious experiences*. But since those reasons are limited, in explaining her dispositions to act, it will never be correct to attribute to her any specific large number belief, such as the belief *that exactly 1,920,762,973 grains of sand are there*. For the same reason, no matter what she does, it will be never correct to assign to prelinguistic Sally beliefs with content (a), (b) or (c).

Given modified interpretationism, then, the only way for Sally to believe sophisticated contents like (a), (b) and (c) is outer-language-first as opposed to mind-first. She must learn an outer language. There will be some story about how the sentences of her outer language come to express (a), (b) and (c) that does not appeal to an explanatorily prior ability to believe those contents. And then she can believe those contents by 'understanding' and 'accepting' those outer sentences. Surely this direction of explanation is possible. Why not think it is sometimes actual?

Modified interpretationism, then, may better fit with a few features of our concept of belief than Williams's variety. Also, notice that, when it comes to sophisticated beliefs, it is quite similar to his actual view. The only difference is that while Williams holds that our believing (a), (b) and (c) is grounded in our 'accepting' hidden *inner* sentences expressing those contents, the modified view holds that it is grounded in our 'accepting' *outer* sentences expressing them. Therefore, in moving to modified interpretationism, he could retain his main ideas: for instance, his rationality maximization theory of how terms and sentences get their contents, and his functionalist theory of what it is for a sentence to be 'accepted' (in the 'belief-box'). He could simply apply them directly to outer sentences.

Modified interpretationism rejects Williams's inner isomorphism constraint. So how does it avoid the intuitively mistaken verdict that 'Blockhead' has beliefs and desires (33–35)? The answer is that on modified interpretationism understanding and therefore belief require source intentionality involving *consciousness* and causal-inferential connections to such source intentionality. Blockhead is an unconscious robot that fails to satisfy this internal constraint.[17]

6. *How does rationality maximization apply to 'plus' and 'game'?*

Finally, a general question. Williams ingeniously develops a rationality-maximization theory of content-determination for some of Sally's inner LOT terms: logical terms like 'and', explanatory terms like 'motion' and moral terms like 'wrong'. How does the view generalize to other terms?[18]

For example, consider Sally's LOT terms 'plus' and 'is a game'. On Williams's rationality maximization view, what are the basic underlying non-representational (inferential or dispositional) facts that determine what they denote? And, for these two terms, what are the relevant basic principles of substantive rationality, such that the correct assignment of denotations to them is the one that maximizes Sally's conformity to those principles? Is rationality really part of the explanation of how 'is a game' comes to express some messy property *P*?

And how does Williams solve underdetermination worries here? What makes it the case that Sally's LOT term 'plus' denotes the plus-function rather than the *quus$_g$-function* – an arithmetical function like the plus-function except that it gives deviant results for a few specific numbers in the googolplex range too large for Sally to compute? And what makes it the case that 'is a game' expresses property *P* rather than property *P**, where *P** has exactly the same extension as *P* across all worlds, except that *P** goes somewhat deviant relative to a single, extremely remote world satisfying complete description *D* (where *D* is too long and alien for Sally to understand)? Sally's actual finite dispositions do not distinguish between these interpretations; she cannot even consider the specific remote cases. And idealization faces well-known problems (Boghossian 2015). So would Williams appeal to 'naturalness' to favour the straight interpretations over the bent ones here (Pautz 2021a: 299–300)? If so, how might this be derived from his general rationality maximization view?[19]

17 See Chalmers 2012: 467, Mendelovici 2018 and Pautz 2021a for this constraint.

18 The same question arises if we directly apply the theory to outer language, as suggested in §5.

19 In the case of *explanatory terms*, Williams (64) derives a naturalness constraint on interpretation from his rationality maximization view (*via* a connection between naturalness and IBE rationality). But 'plus' and 'is a game' are not explanatory concepts, so here it is unclear how the derivation would go.

Brown University
Providence
RI 02903
USA
adam.pautz@gmail.com

References

Boghossian, P. 2015. Is (determinate) meaning a naturalistic phenomenon? In *Meaning without Representation*, eds. Tebben, Gross and Williams. Oxford: Oxford University Press.

Chalmers, D. 2012. *Constructing the World*. Oxford: Oxford University Press.

Dennett, D. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.

Mendelovici, A. 2018. *The Phenomenal Basis of Intentionality*. Oxford: Oxford University Press.

Pautz, A. 2013. Does phenomenology ground mental content? In *Phenomenal Intentionality*, ed. U. Kriegel. Oxford: Oxford University Press.

Pautz, A. 2021a. Consciousness meets Lewisian interpretation theory: a multistage account of intentionality. In *Oxford Studies in the Philosophy of Mind*, vol. 1, ed. U. Kriegel, 263–314. Oxford: Oxford University Press.

Pautz, A. 2021b. Review of The Metaphysics of Representation. *Mind*.

Williams, R. 2020. *The Metaphysics of Representation*. Oxford: Oxford University Press.