How Wrong Can You Be?*

How badly can you go wrong at one thing while succeeding at another? This paper pursues a version of this question that is provoked by Chapter 3 of Williams's *The Metaphysics of Representation*: How badly can you go wrong at understanding why ordinary things in the world behave as they do, while succeeding at locking onto those things as objects of thought? Williams's discussion entails that success along these two dimensions goes hand in hand. It entails that to count as thinking about ordinary things, a subject must also count as being sufficiently right about why they behave as they do. I think this is a wrong result.

The paper is structured as follows. §1 summarises the relevant part of Williams's Chapter 3. §2 explains how this part of Williams's proposal generates the result I have indicated, and why I think this result is bad. §3 considers some possible lines of reply. §4 attempts a more general diagnosis.

§1 Williams on the permutation challenge

Chapter 3 of *The Metaphysics of Representation* offers (among other things) a solution to an old puzzle for people theorising about cognitive representation of ordinary things. Here is a toy version of this puzzle which will suffice for present purposes. [1]

Take a mundane situation in which we would like to be able to say that a subject is having a thought that attributes a commonplace property to an ordinary object. For example, suppose that subject S is looking at a green ball in a situation involving no causal or perceptual funny business, and thinking a thought she would express by saying 'That is green.' We would like – from the theorist's point of view – to make claims like this:

i) S's singular concept <that> picks out *that ball* (the one she is currently pointing at).
ii) S's general concept <green> picks out the property of *being green*.
iii) S's belief <That is green> is true iff *that ball is green.*

Now consider a permutation of this initial (partial) interpretation of S (Williams calls this a 'twisted' interpretation):

i*) S's singular concept <that> picks out *o*, where (*o* is an object which exists iff the ball does, and which, in every situation in which it exists, stands in relation R to the ball).
ii*) S's general concept <green> picks out the property of being *F,* where a thing that stands in R to the ball is *F* iff the ball itself is green.
iii*) S's belief <That is green> is true iff *o* is *F.*

And now consider what reason there might be to prefer the *that ball is green* interpretation of the S's <That is green> belief to the *o is F* interpretation. The interpretations bring the beliefs in as true relative to exactly the same possible ways the world might be: the world is such that the ball is green iff it is such that *o* is *F*. So if we want to keep (i) – (iii) over (i*) – (iii*) we have a problem: to find reasons for the preference, given that the interpretations demand the same things of the world if the belief is to be true.

---

[1] Compare Williams pp. 60-61

Williams's solution to this puzzle provides a worked example of his recipe for generating assignments of representational contents. The recipe is built around Williams's Substantive Radical Interpretation constraint on such assignments:

Substantive Radical Interpretation (hereafter 'SRI') The correct interpretation is the interpretation which best rationalises the subject's dispositions to act in the light of her experiences (where 'rationalise' is to be read 'substantively', so that the correct interpretation comes out as the one that maximises the extent to which the subject is responsive to genuine reasons). [26]

SRI motivates the following four-step recipe for the assignment of representational contents to concepts (construed as vehicles of representation).[2] Firstly, isolate a core of behavioural dispositions that may be regarded as constitutive of operation with the concept (in Williams's terms, these dispositions constitute the concept's 'conceptual role'). Secondly, uncover constraints on substantive rationality for this general class of concept-deployment (in Williams's terms, these are 'first order' constraints on what it takes for a subject's deployments of this kind of concept to be in normative good order). Thirdly, argue that, given these constraints, the interpretation that best rationalises the core behaviours will treat the concept as representing…X. Finally, conclude that, by SRI, the correct interpretation will treat the concept as representing X.

Williams joins many others in taking it that, for the concepts associated with observational predicates like 'green' and 'in motion', the core behavioural dispositions include dispositions to treat the concept as applicable in response to an appropriate kind of perceptual experience [59]. He also takes it that, for us, the 'core', conceptual-role-constituting behavioural dispositions for observational property concepts include dispositions to formulate explanations, and to use these explanations to arrive at beliefs that outrun what we observe [62, 64]. This is the first-stage of the SRI-based story for the concepts associated with observational predicates.

At the second stage, Williams derives two constraints on rationalising interpretation for the concepts associated with these behavioural dispositions[3]:

Constraint 1 The rationality-maximising interpretation will maximise the goodness of the subject's explanations (the extent to which connections she treats as explanatory actually do explain the phenomena she observes).

Constraint 2 The interpretation that maximises the goodness of the subject's explanations will maximise the extent to which her observational property concepts are assigned what are in fact potent properties (properties in virtue of the possession of which things behave as they do).

Given these materials, Williams's solution to the permutation puzzle unfolds as follows[4]:

1 The correct interpretation of subject S is the one that best rationalises her dispositions to apply observational concepts in response to experience; deploy these concepts in explanation; and use

---

[2] The recipe is introduced by way of a walk-through for the case of the concept of conjunction at pp 38 – 44, and deployed at many points in the book.
[3] I base these constraints on the discussion at pp. 63-64
[4] See p. 64.

these explanations to arrive at further beliefs. [SRI applied to the case of our observational property concepts.]

2 The interpretation that best rationalises the dispositions at 1 maximises the goodness of the subject's observational-property-deploying explanations. [*Constraint 1*]

3 The interpretation that maximises the goodness of these explanations maximises the extent to which observational property concepts are assigned what are in fact potent properties. [*Constraint 2*]

So
4 The correct interpretation maximises the extent to which S's observational property concepts are assigned potent properties. [From 1 – 3]

Now add a pair of metaphysical assumptions:
5 *Being green* is (as a matter of metaphysical fact) a potent property; *being F* is not.

4 and 5 entail
6 The correct interpretation assigns *being green* to S's concept <green> rather than *being F*.

So
7 In a case where the subject is disposed to believe <That is green> in response to experience of a green ordinary thing, the correct interpretation treats this belief as a belief about the thing experienced. [From 6, given SRI]

Note that, though the 'twisted' (i*) – (iii*) interpretation deviates in its assignments to both the subject's singular concepts and her general concepts, the solution to the problem – the account of why the initial interpretation is to be preferred –is running off the 'property' side of things. The solution works by feeding the property-side assumptions at 5 into the account of correctness of assignment of contents to observational property concepts at 4 to give the conclusion that S's concept <green> refers to *being green* rather than *being F* – this is 6. With the property-side work done, we can use the fact that S's experience gives her at least some reason to ascribe this property to the initial interpretation's value for <that>, and no reason to ascribe this property to the twisted interpretation's value for <that> to generate the object-side result at 7. (Compare Williams's way of putting the permutation challenge on p. 54: the twisted interpretation is generated by putting in two twists for each atomic thought, one on the interpretation of the singular concept and one on the interpretation of the general concept. Williams's solution to the challenge gives us a reason to resist the twist to interpretation of the general concept, and when we untwist the interpretation of the general concept, the singular concept untwists along with it.)

§2 The challenge from Macroscopic Maxwell's Demon

The previous section explained Williams's solution to the permutation puzzle. I shall now bring out an aspect of this solution that I think should at least give us pause for thought. Here is a scenario to get us started:

Macroscopic Maxwell's Demon Sally is an adult who has lived from birth in an environment where the course of macroscopic observable events is controlled by a hidden demon. The demon's interventions are completely unsuspected by Sally, though their results are ubiquitous in her experience of the macroscopic world. To keep things simple, suppose that the kinds of events available for observation by Sally are severely restricted. Balls of various colours come in from the left. The demon hits them with its tennis racquet, sending them off in various directions at various speeds. Sally sees the balls moving around, but not the demon's interventions. The ways the demon hits the balls so far have involved sufficient regularity that Sally has been led to form generalisations like 'Red ones turn upwards'; 'An orange one's angle of deviation from its initial path depends on the time interval since the last green one'; and so on. She has refined these generalisations against her experience, and developed a theory which she uses to predict, and treats as explaining, the behaviours of balls of various colours. In fact, these behaviours depend on how the demon swings its racquet. And the demon swings its racquet according to its own whims: the colours of the balls play no role in determining the paths they will take.

Consider Sally watching a red ball approach. <Because it's red, it's going to turn upwards>, she thinks. And ask yourself this question: is Sally's thought *about* the ball. I think the answer to this question is 'Yes', and would need quite a lot of convincing to the contrary. I conjecture that that's probably your initial reaction too. It seems that Sally is thinking about the ball. Whatever has gone wrong for her has gone wrong with respect to her grasp of why things like the ball behave as they do, and this thing-that-has-gone-wrong has left her capacity to think thoughts about the ball intact. However, it is hard to see how Williams's framework can deliver this result.

To see why this is, we must first say a little more about *Constraint 2* (Premiss 3 in Williams's solution to the permutation puzzle). As it stands, *Constraint 2* states a necessary condition connecting goodness of explanation and potent properties. But one apparent lesson of the Demon case is that whether potency of properties transmits to goodness of explanation depends not just on whether the explanation deploys property concepts which are concepts of potent properties, but whether the explanation fits these properties together in appropriate ways. If this apparent lesson is right, *Constraint 2* must be modified accordingly:

Constraint 2 (modified) The interpretation that maximises the goodness of the subject's explanations will maximise the extent to which the subject is treated as combining representations of what are in fact potent properties into explanatory beliefs that are appropriate to the ways these properties in fact determine the behaviour of objects that have them.

Now, if *Constraint 2* is modified in this way, we can see that Williams's solution to the permutation puzzle at least does not *deliver* the verdict that I have suggested intuition demands in the Demon case. For in the Demon case, Sally's explanations are wildly off relative to the actual causal potencies of the colour and motion properties. They are so wildly off that it is tempting to say that her deployments of observational concepts do not make it onto the 'goodness of explanation' scale. But if Sally's deployments of observational concepts do not make it onto the goodness of explanation scale, Williams's proposal gives us no reason to say that she is thinking about the ball and ascribing a colour to it, rather than ascribing a colour* to an object assigned to the ball under a permutation.

And there is what looks like a reasonably good argument for a stronger claim. Williams's story not only fails to *deliver* the verdict that Sally is thinking about the ball; it is in fact *inconsistent* with this verdict:

1 The correct interpretation treats Sally's <That is red> thought as about the ball. [Assumption]

2 The correct interpretation treats Sally's <That is red> thought as ascribing the property *being red* rather than the property *F*. (From 1 and 'two twists' structure of the permutation challenge.)

3 An interpretation is correct only insofar as it maximises the potency of the subject's explanations (treats the subject as formulating explanations that in fact reflect potent relations between potent properties).

4 The correct interpretation treats Sally's deployments of <is red> in a way that maximises the potency of her *because it is red* explanations. [From 2, 3]

But
5 Sally's explanations are too wide of the mark to count as reflecting the potency of the properties with which she is in observational contact, so too wide of the mark to enable a genuinely potency-maximising treatment of her *because it is red* explanations. Contradiction.

So we have arrived at the bad result for Williams's framework that I want to bring out. In this framework, a subject counts as thinking about the ordinary things with which her perceptual experience puts her in contact only if she counts as at least doing reasonably well at explaining why these things behave as they do.

To avoid misunderstanding, I should stress that the challenge to Williams that I have raised in this section is not a sceptical challenge. The question I raised about Sally was not 'How do we know we're not in Sally's situation?' It was 'Given her situation, should we say that Sally is thinking about the ball?' I have suggested that the answer to this question should be 'Yes', and shown that, at least on the face of things, Williams's framework commits him to saying 'No'. (There are issues not unrelated to scepticism lurking nearby, as is about to emerge.)

§3 Potential replies

How might Williams respond while keeping his solution to the permutation puzzle? Here are four options, with preliminary discussion.

First option Reject 1 in the 1 – 5 argument from the previous section, denying that Sally's <That is red> thought is *about* the ball.

This move would evade the challenge as I have stated it. But, obviously, the Demon scenario is just one among many structurally similar such cases. If this first tactic is to dispose of the puzzle-raising scenario for good and all, the suggestion must be that no case with this stricture is one where the subject counts as thinking about the objects she is experiencing. And it is here that issues not unrelated to scepticism arise.

To see how, let us first allow that it is not a genuine epistemic possibility that we are in Sally's situation. Her predicament is sufficiently like a silly 'sceptical hypothesis' scenario that, as far as epistemic possibility is concerned, it can be set aside by whatever is our preferred response to external-world scepticism. However, it is a straightforward observation that there have been points in our collective history when our situation has been Sally-like in the following respect: we have have been (as it turned out) wildly mistaken as to why things or individuals behave as they do. We *thought* that all material things are made of water; that objects fall to the

ground because each kind of thing (earth, water, air, fire) has its natural place; that there is no action at a distance; that matter is spatially dense; that mass is one thing and energy another.

If we push this first response on Williams's behalf, we have a choice between two options with respect to subjects who 'explain' the behaviours of ordinary things in ways that we now know to be very wrong. The first is to say that these people do count, from our point of view, as making it onto the goodness of explanation scale.The second is to say that, from our point of view, the subjects do not count as thinking about ordinary things.

To uphold the first response, we would need to argue that one or other of these moves is always going to be defensible. And the problem is not just about what it is coherent for us to say about our now-known-to-be-wrong forbears. It is also about what we are entitled to say about ourselves. Given the history of massive shifts in our understandings of why ordinary objects behave as they do, it looks like lack of historical humility simply to *claim to know* that our everyday explanations capture real potencies in the properties of ordinary things. Someone making this first response is allowing that if we make what looks like the right, non-hubristic move – allowing that we might just be wrong about why things really behave as they do – we lose our right to claim to know that we are thinking about these things at all.

Second option Reject 2, denying that an 'untwisted' treatment of <that> forces an untwisted treatment of <red>.

If this is the response, Williams owes us an account of why the 'untwisting' of interpretations works in one direction but not the other. The solution to the permutation puzzle sketched in §2 takes it that when we untwist the interpretation of general concepts, the twist with respect to singular concepts unwinds too. If 2 is to be denied, we need an account of why this does not also apply the other way around. Alternatively, Williams might reject 2 and abandon the 'untwisting general concepts untwists the singular concepts' part of his solution to the permutation puzzle. But in that case he will owe us an alternative account of why our ordinary singular concepts represent ordinary objects.

Third option Deny that 2, and 3 entail 4. The suggestion here would have to be that there are constraints on genuine essayed explanations which are not met by Sally's deployments of <red>. If Sally is not deploying <red> in genuine explanations, 2 and 3 can be granted, but 4 does not follow.

Someone making this move will need an account of the difference between what Sally is doing and genuine attempts at explanation which does not undermine the claim that our own deployments of observational concepts are genuinely explanatory. (The suggestion might be that only some of our observational property concepts are deployed by us in genuine explanations; that 'untwisting' with respect to these concepts untwists the interpretation of ordinary perception-based singular <that>; and that the untwisted interpretation of ordinary <that> itself untwists the interpretations for non-explanatory observational concepts. Perhaps primary property concepts untwist first; this untwists the interpretation of <that>; and secondary observational concepts come last, with no requirement that our deployments of them are genuinely explanatory. But all of this would require both an account of genuine explanatory deployment of an observational concept, and a reason to think that there can be no Sally-type scenario in which the subject is essaying genuine explanations.)

Fourth option Reject 5, arguing that Sally's deployments of observational property concepts do make it onto the goodness-of-explanation scale. This will require tinkering with *Constraint 2*. I have argued that if there is a connection between potency of properties and goodness of explanation it cannot be just that the interpretation that maximises goodness of explanation maximises the extent to which the property concepts the subject treats as explanatory are assigned potent properties. This would give us no reason to prefer an interpretation that treats a subject as explaining the course of her experience in ways that do capture why things behave as they do to one that treats the subject's explanations as dealing in genuinely potent properties, but scrambling how these properties relate to one another. If 5 is to be rejected, the suggestion will have to be that there is a defensible version of *Constraint 2* which is met in Sally's case, and which nevertheless recognises that explanations must relate potent properties in potency-transmitting ways.

Let me summarise the discussion so far. §2 laid out Williams's solution to the permutation challenge. §3 used the Demon case to argue that this solution commits Williams to what I suggested is an implausible claim: the claim that success at locking onto ordinary things as objects of thought depends on sufficient success in explaining the behaviours of these things in terms of the actual causal potentialities of actually potent properties. This section has considered four potential replies.

Much more could be said about the feasibility of each of these options. But rather than speculating about which might be the best bet for someone wanting to keep Williams's response to the permutation puzzle, I shall close by attempting to generalise the challenge I have raised.

§4 The interpretationist malaise

So far I have focussed on one specific application of Williams's interpretationist framework. But the discussion of this application points towards a general objection that I shall now try to bring out.

Let me start by going back to Williams's SRI:

Substantive Radical Interpretation (SRI) The correct interpretation maximises the extent to which the subject is treated as responding to reasons provided by her experience.

Now, why are we supposed to be tempted by SRI in the first place? I take it that the idea is something like this. Suppose we start with the claim that a right account of the functional roles of mental states treats beliefs as, in general, rationalised by experience and rationalising action given desires. Then we have at least the following 'content and rationality' constraint on the practice of assigning representational contents to beliefs:

Content and Rationality To count as engaging in this practice, we must go about things in a way that, as far as possible, avoids assignments to the subject of beliefs that are not rationalised by her experience and/or do not rationalise her actions.

Why does the claim about functional role entail *Content and Rationality*? Because given the functional role of belief, a theorist who violates this constraint is not genuinely in the business of trying to work out what the subject *believes* at all. But an interpretational practice which

conforms to this constraint *just is* one that maximises rationality. So if the initial claim about functional role is accepted, it brings with it the basic interpretationist view of correctness of assignment of beliefs and belief contents: the correct assignment maximises the extent to which the subject is treated as forming beliefs that are rationalised by her experience, and rationalise her actions given her desires.

I take this to be a standard line of thought in favour of a general interpretationist approach. SRI results if we add that the relevant notion of rationalisation is *substantive* rationalisation (reason responsiveness).

In Chapter 2, Williams gives us an argument for this last step. The contrast is with the suggestion that the correct interpretation might need to maximise only 'structural' rationality – the extent to which the subject's beliefs 'pattern' in a way that is appropriate given her experience [14]. Williams's point is that the 'substantive' gloss is required to yield a criterion for correctness of interpretations that will rule out 'obviously wrong interpretations' [29]. For example, consider interpretations I and I* which match with respect to what they treat the subject as believing in direct response to her experience, but differ radically with respect to what they treat her as believing about the world beyond her experience: I treats the subject as (in intuitive terms) believing that the world is still there when she is not experiencing it; I* treats S as committed to the existence of only the spatial and temporal parts of the external world that she experiences. Suppose that the only constraint on correct interpretation is that we maximise the extent to which the subject is treated as forming beliefs that are prompted by her experiences and that conform to various internal structural constraints. Then I and I* will be equally correct.

Now, Williams takes it that the obviously wrong interpretations are…obviously wrong, and argues for SRI on the ground that a right interpretationist framework must entail the wrongness of obviously wrong interpretations. But given the role played by the appeal to the functional role of belief in the standard line of thought, there is something else we can say about why, if you are going to be an interpretationist at all, you should join Williams in moving to SRI. Let us grant that we can fence off some set of requirements for the rationality of a subject's beliefs which are 'structural' requirements (that the beliefs form a pattern which licenses the claim that they are prompted by her experiences; that the beliefs meet various internal consistency and coherence conditions). Suppose Williams is right that there are requirements on being rational that go beyond the mere structural requirements – call them 'substantive' requirements. Then as long as there are interpretations that are ruled out by substantive + structural requirements but not by structural requirements alone, assigning contents to the subject's beliefs in a way that maximises her conformity to the structural requirements is not assigning contents to her beliefs in a way that maximises the extent to which she is treated as forming beliefs that are rational given her experiences. But in that case, given the initial point about functional role, 'interpreting' in a way that maximises only conformity to structural requirements is not genuinely treating the subject as forming and holding beliefs at all. So if the the initial point about functional role gets us as far as the *Content and Rationality* constraint, it gets us all the way to SRI: to follow the standard line of thought as far as *Content and Rationality* but refuse to join Williams in endorsing SRI is incoherent.

Against this background, the general challenge to Williams's framework that I want to bring out can be put like this.

1 A right account of the functional roles of mental states treats beliefs as, in general, rationalised by experience and rationalising action given desires. [Supposition.]

2 The supposition at 1 generates SRI – the claim that the correct interpretation of a subject is the interpretation that maximises the extent to which the subject is treated as responding to reasons provided by her experience. [Conclusion of this section so far.]

3 SRI entails that factors that block treating a subject as genuinely reason-responsive with respect to a given part of her mental life also block treating this part of her mental life as involving the formation of beliefs which succeed in being about anything. [General version of the aspect of the SRI-based framework uncovered by discussion of the Demon case.]

4 Factors that block reason-responsiveness may leave aboutness intact. [The current author's visceral conviction, intended to be advertised by the Demon case, and bolstered to some extent by the discussion of the first option in §3.]

1 – 4 are inconsistent, so one of them has to go. The discussion of earlier sections was centred on 3 and 4. But my own view, for what it is worth, is that the real guilty party is 1. I think Williams is just right with respect to much of what he says about what the most plausible interpretationist view might look like: in many respects, *The Metaphysics of Representation* is a master-class in how to be an interpretationist, if you want to be one. Some of the things about which I think he is right are folded into 2 and 3. But if 1, 2, and 3 are right, 4 is wrong. And this cannot be. We need to say something else about the functional role of belief.

References

Williams, J. R. G *The Metaphysics of Representation*. Oxford, Oxford University Press. 2019.

---

* Thanks to Robbie, to everybody who was at the Author Meets Critics session where the three critics in this symposium presented initial versions of our comments (at that 2020 Central Division Meetings of the American Philosophical Association); to Phil Kremer; and to members of the graduate class that Phil and I co-taught at the University of Toronto in 2021.