Article:
Williams, JRG orcid.org/0000-0003-4831-2954 (2021) Reply to Critics. Analysis, 81 (3). pp. 536-548. ISSN 0003-2638

https://doi.org/10.1093/analys/anab049

Reply to critics.
JRG Williams

Dickie, Pautz and Chalmers provide three probing, creative critiques of *The Metaphysics of Representation* (Williams 2020, henceforth: MR). I select some highlights to discuss, regretfully passing over much of interest.[1]

Reply to Dickie.

Consider a twisted interpretation of Sally's concepts, assigning wildly wrong denotations to each, but constructed so that the twists cancel out in such a way that the coarse-grained truth-conditions of the thoughts always match the correct interpretation.[2] In virtue of what are such twisted interpretations incorrect?

MR ch.3 says: twisted interpretations of our concepts assign bizarre, unnatural denotations to the general concepts we use in framing the hypotheses that make sense of our observations. But such properties are not fit to feature in genuine explanations—they are "impotent". Hence the twisted interpretation paints us as less epistemically reason-responsive than the ordinary, non-twisted interpretation. Against the backdrop of an interpretationism that identifies the correct interpretation of an agent as the most reason-responsive one, this explains why the twisted interpretation is incorrect.

Dickie worries that this grounds determinate reference only in epistemically friendly possible scenarios. She invites us to consider a possible scenario where ordinary properties (in particular, colour) appear potent, but where appearances are deceptive. Sally sees local patterns that would fit with a lawlike generalization <red balls turn upwards>, but the true mechanics in the situation involves a tennis-racquet wielding demon, changing the direction of the balls by whim. Dickie suggests that in this world, we do not make Sally more reason-responsive by assigning *redness* rather than an impotent twisted variant as the denotation of her concept *red*, since (in this scenario) each is as impotent as the other. If similar tricks can be pulled simultaneously for other general concepts, then I lose my stated ground for attributing determinate reference to Sally. Against this, Dickie urges, that Sally's perceptual demonstrative thought, tracking the ball and accurately attributing properties to it, determinately refers to that very ball (for Dickie's own story, see her 2015).

I will here defend the specific way of anchoring determinate reference I discuss in the book and which Dickie here puts under the microscope. This won't address her challenge in full generality, that "factors that block reason-responsiveness may leave aboutness intact", as I consider only whether the specific case she constructs witnesses the truth of that more general claim. In return, I won't try to sneak out of her challenge by appealing to factors other than explanatory potency, nor will I question whether "similar tricks can be pulled simultaneously" for all concepts.

Here's how I'll do it: I will set out a description of what's going on in the demon world that is a rival to Dickie's. The rival description makes it look like my explanation of determinacy will still go through, so that there can be peace between the two of us: we will both agree that Sally can determinately refer to the ball she is tracking. I'll then diagnose how our descriptions of the case differ, and what the underlying dispute turns on.

In the demon world Sally tries to infer to the best explanation of her perceptual evidence, forming the (false) belief that <all red balls turn upwards>. She is thereby led into error. Nevertheless, I say, she is justified in forming that belief, and justified because that is indeed the *best* explanation of the data available to her. The error arose not through faulty reasoning, but by good reasoning applied

[1] I owe thanks to my critics not just for these three pieces, but also for earlier versions presented and discussed at the APA, and for their feedback on these issues throughout the last year. Our interactions have been a wonderful, constructive experience for me.

[2] Cf. Davidson (1979), Williams (2008).

to biased data. *That all red balls turn upward* is, on this telling, still a good *putative* explanation of the data (in a way that a putative explanation built out of twisted unnatural properties would not be). On this telling of the story, in the demon world as in ours, the untwisted interpretation makes her more reason-responsive than the untwisted one.

Dickie's modified "Constraint 2" says that explanations, to be good, need to feature potent properties combined in ways that "are appropriate to the ways these properties in fact determine the behaviour of objects that have them". She says Sally's putative explanation is *not* the best available because within it properties are not combined in a way that reflect the facts in the demon world. I say: mirroring of the behaviour-determining facts is only required for *true* explanation, not for *good* explanation. A good (putative) explanation in the relevant sense needs to built out of potent properties, predict the data, and be as simple as possible, but it needn't be accurate.

Dickie's underlying challenge (premise 5 in her argument) is that Sally's explanation in the demon world is too "wide of the mark to count as reflecting the potency" of properties. There is still mileage in this thought. Rather than articulating it as a demand that good explanations mirror the behaviour-determining facts, think of it as the constraint that good explanations are sensitive to *whether* its ingredient properties determine behaviour. Dickie might say: redness counts as *potent* in a world where it features in lawlike regularities, but is *impotent* in the demon world where it is explanatorily epiphenomenal. For this to be the case, potency must be a contingent quality of properties. The consequence: in the demon world where colours are epiphenomenal, Sally's putative explanation isn't even built out of the right ingredients to count as a good putative explanation of the data. The contingency thesis goes against the thesis that degree of potency is identical to degree of naturalness (assuming, as standard, that naturalness is non-contingent). But if the above is correct, so much the worse for that speculative thesis!

Whether Sally's IBE in the demon world is "too wide of the mark" to lead to justified belief turns on this underlying issue about the contingency of potency. This is the real underlying dispute. How might it be resolved? The proper place is foundational epistemology, where we ask the question: what makes a given belief-forming method justification-conferring? Let potency* be a non-contingent characteristic of properties (e.g. naturalness). Let potency** be a contingent characteristic of properties, possessed by P at w only if P figures in true explanations at w. Redness, since it is an explanatory idler in the demon world, may be potent*, but it is not potent** at that world. Let IBE* be a belief-forming method: to believe hypotheses consistent with our data on the basis of their simplicity and potency. Let IBE** be the same procedure, except substituting potency* for potency in the ranking. The real question between Dickie and myself is: are beliefs formed by IBE* justified, for Sally in demon world, or is it only beliefs formed by IBE** that are justified? Putting the issue in this way enables us to get beyond terminological disputes about what the technical term "potency" means, and to substantial issues about what fills the epistemic role associated with it.

I am attracted to a particular way of tackling questions like this: the meta-epistemic characterization of epistemic warrant of (Plantinga, 1993). Framed in terms of justification, this says the following: for a belief to be justified, it must be the output of a part of a properly functioning belief-forming system that is *aimed at* forming true beliefs, and *reliably succeeds* in producing true beliefs in normal circumstances.

I assume that in epistemically friendly worlds, where potent* properties figure in true explanations, IBE* and IBE** both pass muster by Plantingan lights. If either were written into a belief-forming mechanism aimed at true belief, then each would be equally and sufficiently reliable, in epistemically friendly worlds. After all, in epistemically friendly worlds, the potent* and potent** properties coincide, so the two mechanisms do not differ. When we turn to the epistemically hostile environment of the demon world, however, IBE* will be biased towards putative explanations which do not reflect the character of the world in question; IBE** tailors its resources exactly to the properties that reflect the world's character. Plausibly, IBE* is not successfully aimed at truth in the demon world, but IBE** is. If that is the case, then potency**, not potency*, is the epistemically relevant characteristic of properties. In the demon world, therefore, attributing red rather than its

twisted variant to Sally would secure no epistemically relevant boost. Prima facie Dickie would be vindicated, and I lose.

In response to this threat, I return to the original datum from which I started. I say again that Sally's belief that *all red ones turn upward* is justified, given her data, irrespective of the fact that colour properties are in fact explanatory idlers. To preserve this, I'll need to insist on the thesis that potency* rather than potency** is the epistemically relevant quality of properties, and that IBE* rather than IBE** is the right method of belief formation. But how can I say this and defend the datum, given the theoretical pressure outlined above?

Here's how. Plantingan belief-forming methods need to be successfully aimed at truth *across normal possibilities*. One of the defining characteristics of a world being normal, I say, is that there are no potent explanatory idlers—every potent property does explanatory work within it. This implies the demon world is abnormal, even for those creatures, such as Sally, who live within it. IBE is a reliable form of instance for demon-world Sally *across normal worlds,* no matter what biases it might introduce in abnormal worlds. Sally's beliefs formed by IBE are justified even if they are untrue (biased data doesn't undermine justification). Equally, she is justified even if her belief-forming method is systematically unreliable when it interacts with the objective abnormalities of her environment (abnormal metaphysical structures don't undermine justification).

Reply to Pautz

PI.

In the presence of a tomato, Sally has a conscious experience, with the content: there is a reddish round thing. Pautz says:

> (i) all possible "narrow duplicates" of Sally will have the same phenomenology;
> (ii) all possible creatures with phenomenology matching Sally's will have an experience with that same content;
> (iii) all possible creatures with this kind of experiential content will have pro tanto reason to believe that there is a round thing there.

According to my account of experiential content (MR ch9,10), experiences have this content only when, relative to the system in which it occurs, the experience has the function to be produced by reddish round things. If having a function to be produced by X is grounded in "non-narrow" facts such as a regularity of being produced by X, or a suitable history-of-evolutionary-development, then the teleological properties of experience (including its content) will not be preserved under duplication.

Suppose a narrow duplicate of Sally were created by statistical mechanical fluke in a world containing no (other) living things. (i-iii) implies that this Boltzmann-Sally would have content-laden phenomenology which gives her a reason to believe that there is a reddish round thing in front of her. If Pautz is right, my account is doubly lacking: it is missing an account of Boltzmann-Sally's metasemantically narrow experiential content (by i and ii), and missing an account of a class of experientially-given reasons for her to believe that there is a reddish round thing in front of her (by i-iii).

Accept, for the sake of argument, that phenomenology is narrow, and teleology is wide. I then deny both (ii) and (iii). I say: Boltzmann-Sally experiences do not give her reason-to-believe that there is a round thing there. To back this up, consider the Plantingan characterization of epistemic reasons introduced in the reply to Dickie. On that view, reasons-to-believe arise only as part as a *well functioning* cognitive system where that part *has the purpose of producing* true beliefs. By construction, Boltzmann-Sally's cognitive system lacks this or any other function (unlike Plantinga, I stick with a naturalized teleology!). Since teleological properties are involved in the grounding of the reason-giving relation, narrow duplicates that strip away the basis for teleological properties strip away reasons. I therefore deny there is a narrow source of reasons-for-belief that I fail to accommodate.

I also deny there is any metasemantically narrow kind of experiential content. Pautz describes (ii) as "based on reflection". Metaphysically, says Pautz, phenomenology *just is* representation of a special kind. Interesting as that theory is, I don't find it or (ii) compelling based on reflection alone. I do see some appeal to it! I imagine Boltzmann-Sally: only with effort do I resist attributing representational content. But then again: only with effort do I resist thinking that her reddish experience is *supposed to be* triggered by something red. Best theory of teleology tells me the teleological-attribution is erroneous; if best theory tells me the content-attribution is erroneous too, it is no extra cost (after all, on my telling, they are the same error).

Sadly, I don't have a metaphysics of phenomenology to offer. I could be a sensationalist, embracing (i) and denying (ii). I could be an external representationalist, embracing (ii) and denying (i). Pautz is correct to emphasize that an interpretationist cannot duck these issues forever; for discussion on the merits of each, see (Pautz 2021 b).

PII.

Pautz endorses surface-supervenience (SS): correct interpretation should supervene on surface (non-hidden) facts about an agent, where this includes conscious experience, conscious doings, acceptances of inner or outer sentences, and causal-inferential connections between such states. My interpretationism (MR ch.1) violates SS. My formulations commit me to the following "inner isomorphism constraint" (II): 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.

Here's Pautz's case *pro* SS and *contra* II. Sally is a human whose hidden processes are language-of-thought style pieces of good reasoning. We are to imagine a surface-duplicate of Sally where the hidden facts are very different, where hidden facts work via a connectionist network or by substance dualist magic (that this is possible assumes that the duplicated surface facts are not themselves grounded in hidden facts). Pautz invites us to agree that the beliefs and desires of Sally are also enjoyed by her connectionist and dualist twins.

The case *contra* II requires that there's at least one way of filling in the hidden details of these cases in which Sally's twins have beliefs and desires. This is plausible. Accordingly, I owe a reworking of my account to avoid commitment to inner isomorphism. I pay this debt in the final section of this piece.

The case *pro* SS requires that *every* way of filling in the hidden details of these cases is one where Sally's twins duplicate her beliefs and desires. I think that's implausible. While Pautz is correct that original Blockhead isn't a surface-duplicate of Sally, one way of filling in the hidden details of these cases is the pseudo-Blockhead way: such a creature replicates Sally's conscious experiences, doings and acceptances, but the inner workings (whether implemented by language-of-thought programming, a connectionist network, or dualist magic) implement a massive lookup table. For all the richness of conscious experience and doings, this creature is still as dumb as a toaster. Nobody not already in the grip of SS should agree that this creature enjoys Sally's rich beliefs and desires.

(Will Pautz agree that pseudo-Blockhead is a surface duplicate of Sally? He could question whether the causal-inferential relations between the other surface facts are preserved. Blockhead's processing involves holistic causation between total (conscious) input/output states, mediated by the lookup table, rather than the web of causal-inferential connections between individual belief states of folk theory. The point is moot, pending clarification of what *surface* causal-inferential relations are to be. Suppose a conscious experience E leads to a conscious doing D via a chain of hidden brain states H1,…,Hn. One causal relation between E and D is extrinsic involving the hidden Hi. Neither this, nor the existential generalization thereof (which would require preservation of causal structure) can be what Pautz intends. On the other hand, if

we go for an extremely loose relation whereby E just has to be *somehow* present in D's causal history, then pseudo-Blockhead will count as duplicating these surface causal-inferential facts.)

PIII.
Pautz and I agree that public language is not required for simple beliefs. Pautz thinks a prelinguistic human who has a normal-for-us experiential life would not have beliefs with sophisticated contents involve precise large numbers, disjunctions of mathematical content and bizarre possible scenarios, and complex philosophical theses. Stronger, he thinks that lacking language, and with only normal source intentionality, they *could not in principle* have such beliefs. He thinks this is an datum that requires explanation, citing our putative inability to describe contrary cases. I deny it is a datum. He thinks it's plausible that my account entails such thoughts are prelinguistically thinkable (cf. fn.15). I agree but claim this as a feature, not a bug.

I'll consider the case of large number content (Pautz's other cases will need additional resources, e.g. those of Lewis 1970). My starting point: prelinguistic creatures can in principle have thoughts of arbitrary logical complexity (see ch x). Then: neo-Fregeans (Wright 1983) show us that if such creatures can form a concept *the number of,* implicitly defined by a principle that links numerical identities to purely logical relations of equinumerosity, then every numeral and every arithmetical operation is explicitly definable. Prelinguistic creatures with sophisticated cognition could in principle get from logic to pure mathematical content. (Pautz wonders what local reference-fixing story I could give about what grounds the content of "plus"; neo-Fregean conceptual logicism is where I'd start).

What of applied mathematical content? A prelinguistic creature watches carefully as each sheep enters a pen, retains a record of the event in memory, and is disposed to close the gate when eleven have passed. Underlying this is a Plantinga-acceptable mechanism that transitions from eleven memory traces of distinct sheep, to the judgement that there are eleven sheep in the pen, which combines with a desire that there be exactly eleven sheep in the pen, to cause the action. That desire would itself be modified over time in the light of changes too the flock, showing it to be no hardwired subpersonal tic.

If we agree that a prelinguistic creature can have small-number beliefs and desires, I submit that in analogous circumstances a prelinguistic creature with normal source-intentionality *but supercharged memory* could form precise numerical beliefs with much higher numbers. What's crucial to this is that the creature can transition from millions of memory traces to the matching numerical judgement, which involves supercharged storage capacity and supercharged sensitivity of belief-formation to combinations of states stored therein. The memory *contents* stored can be as simple as they ever were.[3]

Public language matters. In my book I explain how co-expressibility by a single public word imposes interpretative pressures towards shared social content, consistent with mind-first metasemantics (ch.7). Articulating thoughts in public language allows us to enhance our working memory capacity by creating records of beliefs on paper or within other people. Language enables us to pool creativity and processing power by reasoning together. Just as supercharged cognition and memory matter for interpretation, by my lights, so too will linguistically-enabled social supercharging.

Pautz and I both think that language matters to mental content. Pautz's bet is that language matters as a source of borrowed content. I think it matters as supercharger of the abilities that determine content. The right methodology is to develop both pure interpretationism and

[3] Elaborating his modified interpretationism, Pautz says" the correct assignment of beliefs to prelinguistic Sally must be congruent… with the reasons provided by her conscious experiences. But since those reasons are limited, it will never be correct to attribute to her any specific large number belief". I say: complex processing (surface or hidden) of limited experiential contents rationally supports beliefs with complex content.

Pautz's modified version more fully; only when both are fully developed will we have a clear view which is best.

Inspired by Pautz's discussion, here is one way of extending the pure interpretationism of MR. Suppose that *accepting S*, for public sentence S, is a functional state, defined exactly as the language of thought theorists define the relation to mentalese sentences that undergirds belief, belief* (Field 1978). That is: acceptance states are characterized by causal-inferential relations to other acceptance states and inner attitudes*. Next, play the extended mind gambit (Chalmers & Clark 1998): since an attitude* type is a functional type, anything that plays the belief* role is an belief*. Conclusion: we believe* public sentences just as we do mentalese ones. Storage and manipulation of public language is "coupled mental processing". As Pautz observes, MR's mental metasemantics can then run on public or inner words or concepts indifferently. First observation: this would remove the need within the MR framework to posit inner copies of public words; public words themselves will be concepts. Second observation: linguistic metasemantics is now folded into mental metasemantics, rather than being built on top of it Lewis-style, or requiring autonomous grounding Pautz (2021 a)-style. The upshot would be a linguistically-mediated *extended* mind-first metasemantics.

Reply to Chalmers

CI

Chalmers asks: what is the relation between (a sophisticated, flexible, two-tier) inferentialism and my interpretationism?

We are interested in grounding the relation R that holds between concepts and their denotations. In my understanding, the two-tier inferentialist proposes a specific strategy: R is grounded in two steps. First a range of first tier concepts have their denotation fixed by the theory of "source intentionality" (of which more in the next section). Second, for the remaining second tier concepts we construct a relation R*, metaphysically prior to R, which holds between concepts and their "conceptual role", i.e. a set of key inference patterns. For second tier concepts C, R obtains between C and d iff d is the thing which maximizes the validity (or generalized validity, or rationality) of the conceptual role to which C is R*-related.

An analysis of the concept/conceptual-role relation R* is at the heart of this inferentialism. And the analysis must be general: it's no good just having to hand a list of five or six plausible examples of conceptual roles—we want a principled characterization that generates the whole infinite list, for every second-tier concept any thinker could possible have. I have no such general account of conceptual role to offer, and for my own purposes, I do not need one. It is true that in the book I appeal to the "conceptual roles" of concepts that interest me—logical connectives, observational concepts, moral wrongness, etc. I often invite the reader to agree that the conceptual roles I pick out satisfy the following: if we're to maximize the rationality of a subject overall, then we'll be maximizing the rationality of *these* specific inferences for *that* specific concept ("conceptual role determinism"). But I don't need a general notion of conceptual role to formulate the specific theses, and what I say can't be turned into the general notion the true inferentialist needs (I invite the reader to give it a go). I have no recipe up my sleeve for generalizing what I say to the concepts of cat, game or carburettor, let alone the concepts that characterize the thought of octopuses, aliens, group agents and the like.[4]

Some (merely) possible thinkers have a conceptual architecture where each concept comes paired in a specifiable way with a specific, compact, conceptual role, which via some neat Peacockian

[4] Chalmers and Pautz would both like to hear more about what the local reference-fixing story would look like for cat, game or carburettor. The default for an interpretationist is a holistic-non-theory of reference fixing—only independently identifiable local patterns need recapturing. I do offer explanations of *general regularities* in world-concept relations that cover these and many others, such as the potency-magnetism of concepts central to explanation, or social magnetism of those expressible in language.

determination theory determines its denotation. For some possible thinkers there may be some more holistic but still systematically-characterizable conceptual role (e.g. inferential dispositions from a total scrutability-base world description to judgements of the form: x is F—for which see Chalmers 2012). But for you and I my bet would be that while some interesting parts of our own cognition are organized in ways that approximate neat stories of this kind, there's a lot which is much messier.

Interpretationism thus simulates local metasemantic inferentialism—but in a case-by-case way. This can be illuminating even if actual humans don't perfectly fit the stories I tell. First, we might *approximately* embody the conceptual roles specified. Second, it might be *generically* a reason-maximizing interpretation makes the specified inferences rational, but that there are exceptions. Third, this connection might hold only so long as other independent conditions are met. I think that interpretationism will still have a lot to offer here: it can offer approximate, generic, and conditional predictions about what the denotation of the target concept will be. It is less fragile than the local inferentialisms it simulates.

Interpretationism might also be, as Chalmers suggests, an ultimate limiting case of inferentialism,. Let R* relate each concept to the totality of all inferences the subject is disposed to accept, delivering a common holistic conceptual role. Then determine content holistically: R relates a series of concepts to a series of objects iff that assignment maximizes rationality of the inferences in their common conceptual role. If we're also extremely liberal about what counts as an "inference", if we forswear syntactic constraints on its relata, and if we understand rationality as substantive rather than structural, then we reach something I find hard to distinguish from my interpretationist foundational account. Stretched this much, I wonder whether the term "inferentialism" remains a useful taxonomic category, but others may find it so.

CII

Chalmers thinks that a theory of source intentionality is a powerful resource, underutilized. An illustration of this would be if source intentionality already contained the resources to answer the "bubble puzzle" (MR ch.2), without my appeal to substantive rationality.

Suppose there is a language of perception, as well as a language of thought; suppose that it makes sense to talk of inferential relations, and validity, over the combined language of perception-and-thought. Against that backdrop, Chalmers sketches the following mental metasemantics:

(1) The ("spatial") words in Sally's language of perception denote standard spatial relations.
(2) Inferential relations (or inclusion) between Sally's language of perception and her language of thought ground the fact that her spatial concepts, denote standard spatial relations.
(3) The bubble interpretations of ch x assign something other than standard spatial relations to Sally's spatial contents.

(2) and (3) entail that deviant bubble interpretations are incorrect, and (2) requires at most *structural* rationality-maximization, a strictly weaker resource than the one I appeal to. To earn this, we need thicker resources to ground (1): we need a theory of source intentionality that grounds the content of words of the language of perception. In my book, I ground truth-conditional content of whole perceptual states. The former determine the latter, but not vice versa.

The teleosemantic grounds for whole-state content was roughly this: perceptual state s has content p iff s has the function (within the relevant system) to be caused by p. We'd need some sort of extra idea to tease out a word-to-entity relation from these materials, especially since the relata of causation (events, states, propositions, etc) are of a different category from e.g. spatial relations. But we can try! What about:

> w has content S iff w's function, within the perceptual system, is to be produced by some event which, for *some* objects x,y, is the event of x's bearing S to y.

First trouble: let S* a permuted variant of S, so that there's a permutation f of objects such that necessarily xSy iff f(x)S*f(y). Consider the following:

> w's function, within the perceptual system, is to be produced by some event which, for *some* objects x,y, is the event of x's bearing S to y
> iff
> w's function, within the perceptual system, is to be produced by some event which, for *some* objects x,y, is the event of x's bearing S* to y.

This biconditional will be true unless our account of causation/production, or of function, is hyperintentional. I didn't want to assume hyperintentional ingredients in the basis of my story about what grounds representation. Moral: this won't ground determinate reference to spatial relations (over permuted variants) in the language of perception.

Let S^ the deviant reinterpretation ascribed to the agent by a bubble-interpretation, so that necessary, for all x and y within the agent's bubble, xSy iff xS^y. I fear we will find the following holds:

> w's function, within the perceptual system, is to be produced by some event which, for some objects x,y, is the event of x's bearing S to y,
> iff
> w's function, within the perceptual system, is to be produced by some event which, for some objects x,y, is the event of x's bearing S^ to y.

Here xSy and xS^y are not necessarily equivalent, so the worry with S* won't apply directly. Still, I think there's a fair chance the biconditional is true. By construction the S and S^ are locally-equivalent relations, and "locality" can be stretched as far as you like—it could encompass not just the perceptible and manipulable environment of the target of interpretation, but also those of all her ancestors.

At the very least, the issues just flagged show the devil will be in the details of getting a word-level grounding story for perception to work. Someone who was committed to this methodology, of course, will throw resources into the project. Maybe more traditional causal theory of reference ideas will help; maybe the phenomenal intentionality research programme rides to the rescue. But this illustrates the strategic choice we face: to resolve problems with determinacy of reference by goosing up a theory of source intentionality, or to stick with thinner resources, and use the constraining power of substantive rationality to finish the job.

Interpretationist biconditionals, redux.

I intended the general framework of interpretationism to apply to possible thinkers who think with maps rather than sentences. And whatever the prospects of defending the language of thought hypothesis for humans, I did not want to commit to the claim that octopuses, aliens, future general AI, or group agents, are all structured this way.

Chalmers says "Williams appears not to give a state-based foundational story that doesn't involve a language of thought." Fair! Earlier I conceded to Pautz that we shouldn't make it a necessary condition on possible thinkers that they have inner states that map 1-1 to their beliefs and desires. What I said about the basic interpretationist biconditionals (MR ch1) needs fixing, as follows.

I picture a space of possible *interpretations* of agent x*,* each point in the space telling us what x believes and desires at each time. The metaphysician of representation gives an illuminating analysis of what makes a point in this space correct. But what are these interpretations like? Interpretations might map whole *stages* of persons to a belief-desire pair; they might map inner *states* of persons to attitude-content pairings; taking these inner states to have language-like structure, an interpretation might map the "word-like" parts of these states to subpropositional contents such as objects and properties while giving compositional rules for determining the contents of "sentence-like" states. Different cognitive architectures yield further options.

The three options mentioned are not in direct competition. If J is a point in the space of compositional interpretations, J assigns content to the word-like parts of x's inner states and specifies the rules required to thereby determine the content of states as a whole. So modulo an identification of the attitude-type of the states (e.g. by functional role) J fixes a point J* in the space of state-based interpretations. With J* to hand, we have a complete specification of the content of x's explicitly-represented belief and desire states at each time, and modulo a theory of how explicit and implicit attitudes relate (e.g. implicit attitudes identified with dispositions to explicitly-represent), this in turn picks out a point J** in the space of stage-based interpretations.

Languagelike thinkers will need to be interpreted by picking out points in a space of language-like interpretations. Maplike thinkers will need to be interpreted by picking points in a space of map-like interpretations. Whatever fine-grained space we start in, we end up picking out a point in a common, neutral space of stage-based interpretations.

Let S_x pick the finest-grain interpretational space relevant for creatures of x's kind (the space "tailored" for x). I say:

> Necessarily, for all x, x believes that p iff J says that x believes that p, where J is the correct interpretation of x within the tailored interpretational space S_x.[5]

By construction, the tailored interpretational spaces are compositional interpretations for languagelike thinkers, cartographic interpretations for maplike creatures, and so forth. The distinctive interpretationist hypothesis about correct interpretation is as follows:

> Necessarily, for all x, J is the correct interpretation of x within interpretational space S_x iff J the most rationality-maximizing interpretation of x (given x's dispositions to act and evidence) within S_x

The view just described makes sense of what I'm doing in the chapters where I work under the language of thought hypothesis. I am exploring interpretationist metasemantics as it applies within one particular tailored space, the tacit assumption being that it is the one relevant to interpreting you and me. The biconditionals here make room for parallel explorations of the implications of the same general interpretationist metasemantics for creatures with different internal structures. Interpretationism, so construed, does not imply what the inner isomorphism constraint (i.e. that beliefs and desires match up 1-1 with interpretable states).

The hidden facts in our heads still matter, since they fix what tailored interpretational space is relevant for us, and this feature of the metasemantics still blocks Blockhead. Assume Sally's tailored interpretational space is some refinement of a state-based one, and the correct point in this space makes her maximally rational, given her acts and evidence. Blockhead has the same evidence-to-act dispositions. But as I say in (MR p.34) the correct interpretation of Sally "can't be be the one that is true for Blockhead, since Blockhead is not in the same or even isomorphic states". The tailored space for Blockhead doesn't overlap with the tailored space for Sally! A fortiori, the most rationalizing fine-grained interpretation of Sally cannot be the most rationalizing fine-grained interpretation of Blockhead. We need not assume all thinkers share a common fine-grained interpretational space to get this result, and I withdraw any suggestion to the contrary.

Bibliography

Chalmers, David (2012). *Constructing the World*. Oxford University Press.
Clark, Andy & Chalmers, David J. (1998). The extended mind. Analysis 58 (1):7-19.
Davidson, Donald (1979). The Inscrutability of Reference. Southwestern Journal of Philosophy 10 (2):7-19.
Dickie, Imogen (2015). Fixing Reference. Oxford University Press.

[5] J says that x believes that p iff the interpretation J induces in the stage-based interpretational space does so.

Field, Hartry (1978). Mental representation. Erkenntnis 13 (July):9-61.
Lewis, David (1970). How to define theoretical terms. Journal of Philosophy 67 (13):427-446.
Neander, Karen (2017). *A Mark of the Mental: A Defence of Informational Teleosemantics.* Cambridge, USA: MIT Press.
Pautz, Adam (2021 a). Consciousness Meets Lewisian Interpretation Theory: A Multistage Account of Intentionality. In Uriah Kriegel (ed.), Oxford Studies in Philosophy of Mind.
Pautz, Adam (2021 b). *Perception*. *Routledge.*
Plantinga, Alvin (1993). *Warrant and Proper Function*. Oxford University Press.
Williams, J. R. G. (2008). The Price of Inscrutability. Noûs 42 (4):600 - 641.
Williams, J. Robert G. (2020). The Metaphysics of Representation. Oxford University Press.
Williams, J.R.G. (forthcoming). "Commitment problems in the naive theory of belief" forthcoming in *Unstructured Content,* Kindermann, van Elswyk, and Egan (eds), OUP.
Wright, Crispin (1983). *Frege's Conception of Numbers as Objects*. Aberdeen University Press.