

Taking Alien Minds Seriously

Nonstandard Cognition & the Scope of Philosophy

Abstract

Philosophical theories typically aspire to generality, yet they have largely developed from o with human psychology primarily in mind. Theories of justification, agency, and welfare appeal to belief, desire, perception, and other familiar psychological properties and relations (the *standard mentals*). But there is no guarantee that these kinds adequately represent the space of possible minds. Traditional functionalism allows variation in the physical realizers of mental properties while retaining the familiar psychological taxonomy. Drawing on Bealer's non-reductive functionalism, I argue for *narrow mental pluralism*: distinct mental properties may instantiate the same essential functional architecture while differing in their intrinsic natures. I then argue that narrow pluralism undermines the motivation for *Architectural Uniqueness*, the assumption that every possible mind must instantiate the functional architecture of the standard mentals. This motivates *broad mental pluralism*, allowing minds with genuinely different architectures. The implications extend beyond our theory of mind. Narrow pluralism challenges any philosophical theory that depends on the intrinsic character of the standard mentals, while broad pluralism challenges any that presupposes their functional organization. I conclude with two cases studies, involving epistemic justification and agency, and I argue that taking alien minds seriously requires reconsidering the generality of familiar philosophical categories and may demand the development of new concepts.

Keywords: alien minds; functionalism; mental pluralism; non-reductive functionalism; psychological architecture; conceptual engineering; artificial intelligence

Philosophical theories typically aspire to generality. A theory of justification aims to say what justification is for any subject, a theory of agency what it is for any being to act, and a theory of welfare what makes any subject's life go better or worse. Yet these theories are developed largely from our familiarity with human psychology. They appeal to belief, desire, perception, intention, pain, and other familiar psychological properties and relations—what I will call the *standard mentals*. Beginning with these properties is natural, perhaps unavoidable. But there is no guarantee that the distinctions familiar from human psychology adequately represent the space of possible minds.

The traditional functionalist appeal to multiple realizability does not address this concern. It allows variation in the physical realizers of mental properties while retaining the familiar psychological taxonomy. I approach the problem instead through the non-reductive functionalism motivated by Bealer's Self-Consciousness Argument (§1). Within this framework, distinct mental properties may instantiate the same essential functional architecture while differing in their intrinsic natures. I call this *narrow mental pluralism* (§2). But if the standard mentals do not exhaust mentality, there is no evident reason to suppose that their essential functional architecture

exhausts the possibilities for psychology. This motivates *broad mental pluralism*, the possibility of minds with genuinely different functional architectures (§3). The contrary assumption, *Architectural Uniqueness*, threatens to reproduce at the level of psychological architecture the chauvinism traditionally associated with restrictions on physical realization.

The implications extend beyond the philosophy of mind. Narrow pluralism challenges philosophical theories that are articulated in terms of the intrinsic character of the standard mentals, while broad pluralism challenges those that presuppose their characteristic functional organization. As the cases of epistemic justification and agency illustrate (§4), theories developed around familiar psychological kinds may consequently face substantial problems of generalization. Taking alien minds seriously requires us to determine which features of familiar cases are genuinely constitutive of the phenomena under investigation and which reflect only our particular psychology. In some cases, this may require not merely generalizing existing philosophical concepts but developing new ones.

1. Reductive Functionalism & The Self-Consciousness Argument

On the standard functionalist recipe, we begin with a psychological theory $\mathcal{P}$ whose theoretical vocabulary includes predicates for the standard mentals. $\mathcal{P}$ may be assembled from the platitudes of folk psychology, as in analytic functionalism (Lewis 1970, 1972), or from an empirical psychological theory, as in psychofunctionalism (Block 1978). We then replace the mental predicates (or nominalized forms of them¹) with variables yielding a matrix theory, P , that allows us to isolate the characteristic pattern of the standard mentals. Finally, we existentially quantify these variables. On standard reductive functionalism, the quantifiers are restricted to first-order, typically physical, realizer states and the mental properties are defined as higher-order properties.

The SCA turns on the fact that many psychological relations are ineliminably *intensionally non-well-founded* (Bealer 1997, 2001, 2010). A canonical instance of intensional non-well-foundedness is given by the following self-intimation principle:

$\mathcal{J}$. If x believes that p and x introspects, x will believe that x believes that p .

The important feature of $\mathcal{J}$ is that the very psychological property being defined (in this case, believing) can occur in an intensional position as part of the content of that property. This creates a dilemma for the reductive functionalist. Consider two options for generating the Ramsey sentence for $\mathcal{J}$.² On the standard functionalist recipe, we remove

¹ Bealer extends the argument to Lewis's form of functionalism by distinguishing "pain" from the property of "being in pain" (1997, pp. 80–82; 2001, p. 303, n. 8 in response to Tooley 2001). Even granting that "pain" nonrigidly designates whichever first-order state occupies the pain role, the gerundive phrase "being in pain" rigidly designates a property, just as "being the tallest man" designates the same property although "the tallest man" can designate different individuals. On Lewis's proposed analysis, being in pain is the property of having whichever first-order state occupies the pain role, and is therefore a second-order property. Treating "pain" as an occupant-designating expression thus does not avoid the problem, since the relevant embedded predication concerns the property of being in pain. (It is also worth noting that the context here is intensional, so even if r_1 is (contingently) pain, the substitution is problematic.)

² For simplicity, ignoring the rest of $\mathcal{P}$.

all of the theoretical vocabulary from $\mathcal{J}$, replace it with appropriate variables, and bind those variables with quantifiers restricted to first-order physical realizer states:

$$I_R. (\exists r_1, r_2)(\text{If } x \text{ } r_1 \text{ that } p \text{ and } x \text{ } r_2 \text{S, } x \text{ will } r_1 \text{ that } x \text{ } r_1 \text{ that } p)$$

There are two significant concerns with the standard recipe. First, (I_R) requires that there be an intensional, non-well-founded first-order physical relation, r_1 , that has itself as content. But it is far from clear that there are any physical relations that exhibit the appropriate (primitive) fine-grained intensionality, much less the kind of intensional non-well-foundedness required of r_1 .

But putting that aside, the standard functionalist recipe gets the characteristic pattern of the mentals wrong. To see this, we need to distinguish between the occurrence of a psychological predicate, ψ_1 , in its psychological role and the occurrence of that predicate in its role as intentional content. In its role as intentional content for some psychological state ψ_2 , ψ_1 serves as part of the characteristic pattern of ψ_2 ; that is, it needs to function as part of the *definiens*. To illustrate, consider the following:³

$\mathcal{J}$. x believes that pain is to be avoided.

If we were to Ramsify $\mathcal{J}$ we would get the following:

$$J_R. (\exists r_1, r_2)(x \text{ } r_1 \text{S that } r_2 \text{ is to be avoided})$$

Quantifying into the embedded clause distorts the pattern of the mentals that the functional definitions are supposed to isolate. The functional pattern involves a relation to a *specific proposition*. If the Ramsey definition is forced to “swap in” a different proposition, that pattern won’t be preserved. Consequently, in order to get the right functional pattern, the functionalist should leave the embedded predicate untouched. So long as “pain” could be antecedently and independently defined, this would not cause a problem.⁴

But what goes for $\mathcal{J}$, goes for $\mathcal{J}$ as well. If we quantify into the embedded clause we end up distorting the characteristic pattern of belief: It is no part of the characteristic behavior of belief that it stand in any particular relation, much less any essential relation, to a proposition about some first-order realizer state. In other words, since the relation to the belief-involving proposition is itself part of the characteristic psychological pattern being Ramsified, changing the embedded content distorts the functional pattern.

Consequently, just as with $\mathcal{J}$, the correct move would appear to be to leave the embedded predicate untouched. But in cases like $\mathcal{J}$, we cannot do this. If the mental predicate is retained within the embedded content so as to preserve the pattern, the proposed definition makes essential use of the theoretical vocabulary on the right-hand side of the definition. The difficulty is not optional. Psychological theories evidently contain non-well-founded attitudes and this embedding appears to be an essential feature of the intentional organization relevant to complex, intelligent thought. Reductive functionalism is thus forced to assign intensional, non-well-founded properties incompatible roles as both (part of the) *definiens* and simultaneously as the *definiendum*.

³ The example isn’t intended as a genuine candidate for inclusion in the Ramsey definitions.

⁴ See McCullagh (2000) for discussion of stepwise functional definitions; see Bealer (2000, 2001) for related replies. Båve (2017) develops a response using conceptual-role semantics for LOT sentences.

The SCA, however, does not show that functional organization is irrelevant to defining the mental properties. It shows that the standard reductive strategy fails. The obvious question is therefore whether the functional organization can be preserved without the reductive ambition.

2. Non-reductive functionalism and narrow pluralism

Although the SCA is arguably fatal for reductive functionalism, it isn't fatal to functionalism. In response, Bealer proposed to go non-reductive. In non-reductive functional definitions, we remove the restriction on the quantifier to ontologically prior first-order realizers and allow them to range over the mental properties themselves (Bealer 2010, 154–58). Since the definitions are abstracted from the theory of the standard mental properties, the standard mental properties trivially satisfy them. If no other natural, non-Cambridge properties satisfy the resulting definitions, then the definitions uniquely isolate the standard mental properties. Call this the Uniqueness Hypothesis.

There is a plausible thought behind his proposal. The SCA reveals a distinctive intensional non-well-foundedness in psychological organization. Mental relations can figure within the contents of other relations belonging to the same psychological sequence, including, notably, themselves. There are, at least, no obvious familiar physical relations exhibiting primitive intensional non-well-foundedness of this sort. Bealer may therefore be right that this feature is a mark of psychological level properties. If so, a natural sequence instantiating the relevant intensional, non-well-founded structure has a strong claim to being psychological rather than being a merely physical or otherwise non-mental sequence.⁵ But this proposal about the psychological level should be distinguished from the Uniqueness Hypothesis. It is one thing to hold that intensional non-well-foundedness is characteristic of the psychological level; it is another to hold that the standard mental properties are the *only* natural mental properties capable of exhibiting it. The latter claim is, I will argue, mistaken.

Recall that in the non-reductive setting *the mental properties themselves are playing the role of the realizer states*. Familiar counterexamples to functionalism, such as inverted spectra (Block 1978; cf. Shoemaker 1982), tell equally against reductive and non-reductive functional definitions. In the non-reductive setting, however, their upshot is different. Rather than merely showing that the definitions leave something phenomenological out, they support the possibility of distinct mental states. Consider a creature who has a state with the characteristic causes and effects associated with our pain, but whose phenomenal character is nausea-like. Call this “*n*-pain”. Assuming that a pain has its phenomenal character essentially, *n*-pain $\neq$ pain. Intuitively, *n*-pain is no more Cambridge-like than pain itself. But given both its phenomenal character and its functional identity with pain (i.e., is part of a sequence of properties that satisfy the core principles of psychology), there is a strong case that it is a genuine, but alien, mental property.⁶ If so, then we have reason to believe that the Uniqueness Hypothesis is wrong. There are at least two distinct sequences of non-Cambridge mental properties that satisfy $\mathcal{P}$.⁷ And, of course, if a familiar

⁵ This is a slightly different claim than saying that intensional, non-well-foundedness is the “mark of the mental”. Instead, it is a claim that intensional, non-well-foundedness is the mark of the mental *level*.

⁶ By “alien” mental property, I mean here and throughout, a mental property that is not one of the standard mental properties we happen to instantiate.

⁷ If such cases are admitted, they show that the “hard problem” of phenomenal consciousness (Chalmers 1996) does not arise from the physicalist assumptions of traditional functionalism but, instead, directly from functionalism

phenomenal character can occupy an unfamiliar role, nothing in the role excludes a phenomenal character outside our repertoire altogether (Nagel 1974).

The argument to this point suggests that there may be indefinitely many non-Cambridge mental property sequences that satisfy Bealer's definitions. This implies what we might call Weak Narrow Mental Pluralism:

WNMP: There are possible sequences of natural properties, distinct from the standard mentals, that instantiate the same functional architecture and are themselves sequences of mental properties.

WNMP is sufficient to defeat Bealer's non-reductive definitions. But it is a relatively weak thesis. A natural further question is whether any sequence that satisfies the definitions fails to be a mental sequence. Before considering specific candidates, note that two general considerations support a negative answer. First, every admissible sequence must instantiate the intensional non-well-founded structure characteristic of mentality. If that structure is distinctive of the mental level, every admissible sequence contains properties with a strong claim to being mental. Second, every admissible sequence instantiates the organization captured by $\mathcal{P}$, where $\mathcal{P}$ describes the essential functional architecture of core psychology.⁸ Candidate counterexamples can be ordered by their distance from the standard mentals. At each step the question is whether some new feature of the case either prevents the sequence from satisfying $\mathcal{P}$ or gives an independent reason to deny that its occupants are mental. Absent such a feature, these same two considerations apply.

The most obvious place to look for non-mental sequences is to suppose that some admissible sequences are mixed: some of their members are mental while others are not. The most natural candidates for non-mental satisfiers are sequences with non-phenomenal ("zombie") occupants. Two worries arise here. First, the claim that such sequences cannot satisfy $\mathcal{P}$ at all, and second that if they do, they are not mental. Smithies's (2019) work suggests the first. Let *z-pain* be a natural, *non-phenomenal* property proposed as an alternative occupant of the pain-role. A simple one-at-a-time substitution may fail. If standard introspection accesses pain partly by being sensitive to its phenomenal character, replacing pain with *z-pain* while leaving standard introspection fixed could disrupt the relations required by $\mathcal{P}$. Smithies' view of the epistemic role of consciousness provides a version of this sort of objection. Specifically, he argues that blindsighted subjects rationally withhold belief about their blind fields, indicating that the unconscious visual information available to them does not itself supply justification. Merely adding an unprompted disposition to form beliefs, as in super-blindsight, would not supply the missing justification. He connects this epistemic deficiency with introspective accessibility. His proposal is that justifying states must be accessible to introspection, and non-phenomenal visual states lack the requisite access. But even granting his assessment of these cases,⁹ a failure of our familiar introspective capacities to access non-phenomenal perceptual states does not establish that *no* introspective capacity could access them. Smithies' cases retain the standard

itself. That is, functionalism is structurally incapable of capturing phenomenal consciousness, regardless of ontological background assumptions.

⁸ This qualification is necessary (Current Author) and explains the determinate/determinable architecture introduced in fn. 14 below.

⁹ In making this point, I do not wish to concede that human introspection requires phenomenology. I am skeptical of this claim even for the human case.

introspective relation while varying the states to which those capacities would need to provide access. They therefore do not exclude a *coordinated substitution* in which both perception and introspection are replaced by occupants that preserve the relations specified by $\mathcal{P}$. In other words, Smithies' argument shows at most that determinate substitutions can impose constraints on one another, not that non-phenomenal determinate sequences are impossible. Z-pain may require z-introspection which provides access to states in ways that are not sensitive to the presence of phenomenology. And that substitution may in turn require further coordinated substitutions. If a coordinated sequence preserves the relations specified by $\mathcal{P}$, including intensional non-well-foundedness, it is unclear what principled objection there would be counting it as a mental sequence.

The limiting case would be a sequence in which every occupant is non-phenomenal, while the sequence as a whole preserves the relations specified by $\mathcal{P}$. A full zombie sequence of this sort would require not merely non-phenomenal counterparts of pain and introspection, but non-phenomenal occupants of the intentional roles generally. Here a proponent of phenomenal intentionality might object that the substitutions cannot be completed: phenomenal consciousness is essential to the intentionality that the resulting sequence must preserve. However, the present proposal does not attempt to derive intentionality from merely causal relations. In the non-reductive setting, genuine intentional relations themselves occupy the psychological roles, including the intensionally non-well-founded relations highlighted by the SCA. Difficulties in grounding intentionality in nonintentional causal organization therefore do not establish that these relations must be phenomenal. Nor does the fact that familiar instances of intentionality are phenomenally conscious establish that all possible instances must be so. Even perspectival awareness, the subject's occupying a point of view or a center of an "arena of presence", need not be identified with phenomenal consciousness (Current Author).

Eklund (2026; see also Lee 2019) raises a more unfamiliar possibility: occupants that lack phenomenal consciousness but share some positive alternative to it which plays in their mental lives something like the role phenomenal character plays in ours. We need not assume such a feature is possible. The point is only that its being alien to us is no reason, by itself, to exclude it.

If these considerations are sound, the conclusion is not merely that some alternative sequences are mental. No candidate has turned up that satisfies the definitions without being mental, and the two general considerations explain why. This motivates a stronger thesis, Narrow Mental Pluralism, or simply Narrow Pluralism:

NP: Every possible sequence of natural properties that instantiates the same functional architecture as the standard mental sequence is itself a sequence of mental properties.

NP turns the failure of Bealer's definitions into a way forward. If every sequence that satisfies them is mental, the definitions do not pick out the standard mentals, but they do pick out what all such sequences have in common. The idea is to use Ramsification to characterize determinable mental properties, of which the standard mentals are one determinate sequence among others.¹⁰

¹⁰ I use "determinable" and "determinate" in roughly Yablo's (1992) sense: A determinate necessitates its determinable, but not conversely, and having the determinate is a way of having the determinable. Since the relation is a technical one, how best to articulate it remains open (see, e.g., Funkhouser 2023). As far as I can tell, nothing in the argument turns on the choice. Stricter articulations deny that determinates are conjunctions of a determinable

Determinable Functionalism, as we might call it, is a straightforward modification of the traditional definitions. Rather than restricting the quantifier to first order physical realizers (as in reductive functionalism) or leaving them unrestricted (as in Bealer's non-reductive functionalism), we simply restrict the quantifiers to first-order properties. To make the proposal explicit, let $(P(m_1, \dots m_n))$ specify the psychological organization of the standard mentals. We can then define the determinable mental property, M_i as follows:

Determinable Definition: $M_i(x) =_{def} (\exists m_1, \dots m_n)(P(m_1, \dots m_n) \wedge m_i(x))$

Here the quantified properties are the first-order occupants, including the standard mentals. The *definiendum*, $M_i(x)$, is the determinable property corresponding to the i^{th} role. The determinable definitions thus return to their original Lewisian form but are used to define a set of determinable properties instead. What the SCA shows that we must abandon is the restriction to physical realizers, not the restriction to first-order properties. Retaining the latter excludes higher-order sequences, including the determinables themselves, as satisfiers.

Given the SCA, no physical sequence satisfies these definitions; given NP, every sequence that does is a sequence of mental properties. We can therefore characterize the determinate mentals of any given sequence by their determinable together with a differentia. For an arbitrary standard mental:

Definition Schema: $sm_i(x) =_{def} (\exists m_1, \dots m_n)(P(m_1, \dots m_n) \wedge m_i(x) \wedge \Phi_i(m_i))$,

where Φ_i specifies a condition that uniquely differentiates sm_i from other determinates falling under M_i . (e.g., its phenomenological character).¹¹

The failure of the Uniqueness Hypothesis therefore does more than undermine Bealer's definitions. In the non-reductive setting the occupants are themselves mental, so the alternative sequences are not distinct realizations of the standard mentals but distinct mental properties altogether. Determinable functionalism thus moves us from multiple realizability to pluralism at the level of mentality itself.

3. From narrow to broad mental pluralism

Narrow pluralism, as we have just seen, permits a striking range of possible alien minds. Nevertheless, it remains conservative in one important respect—every admissible mental sequence must instantiate the same determinable functional architecture as the standard mentals. Thus, even though the occupants of the relevant roles may differ radically from the standard mentals in their intrinsic natures, they must all instantiate the fixed functional structure obtained by Ramsifying $\mathcal{P}$. In this sense, narrow pluralism allows only architecturally inessential variation between minded beings. This restriction is stronger than it may initially appear. Functional definitions are holistic. The determinable belief-role is partly characterized by its relations to the remaining determinable psychological roles. Those positions are themselves characterized partly by their relations to belief and to one another. Requiring every

with an independent differentia. On those, Φ_i in the Definition Schema should be read as locating sm_i along the dimensions of variation M_i leaves open, not as a genuinely independent conjunct. What the argument requires is simply that determinates necessitate their determinables. Here that follows from $\mathcal{P}$'s specifying only constitutive relations so that each admissible occupant has its place in the architecture essentially.

¹¹ Or, for states such as belief, its location along whatever dimensions the determinable leaves open, such as content grain, cognitive phenomenology, or correctness conditions.

possible mind to instantiate the determinable architecture abstracted from $\mathcal{P}$ therefore does not merely require every mind to possess some collection of states broadly resembling ours. It requires every possible psychology to instantiate the same essential pattern of psychological organization.

When approached from the direction of either reductive functionalism or Bealer's non-reductive alternative, there is an intelligible route to this conclusion. If the standard mental kinds *just are* the mental kinds, one might naturally suppose that every possible psychology must be organized around those kinds. Since the functional definitions involve all and only the essential constitutive relations among them, there is no gap between what is essential to the standard mental kinds and what is essential to mindedness in general. But NP undermines this support. The standard mental kinds do not exhaust mentality. There can be other irreducibly mental properties occupying the very roles from which the determinable architecture was abstracted. Once that is granted, there is no straightforward inference from the claim that $\mathcal{P}$ captures the essential organization of the standard mental kinds to the claim that $\mathcal{P}$ captures the essential organization of mentality *per se*. The functional architecture was obtained by abstracting from one determinate mental sequence, the human sequence. The procedure may reveal genuine essential structure shared by pain and *n*-pain, belief and *alien*-belief, and so on. But it does not follow that the structure is thereby the unique possible architecture of psychology.

These considerations suggest a more general form of mental pluralism:

Broad (Mental) Pluralism (BP): There are possible mental sequences that do not fall under the determinable functional architecture abstracted from the standard mental kinds.

The denial of BP amounts to a surprisingly strong modal thesis:

Architectural Uniqueness (AU): Every possible genuinely mental sequence instantiates the determinable functional architecture abstracted from the standard mental kinds.

AU says, in effect, that our mental organization fixes the boundaries of the psychological level. Every possible mind must contain states organized according to the same constitutive functional patterns we possess. If this were so, the standard architecture of psychology would thus enjoy a modal status quite unlike the merely contingent architectures normally contemplated in other sciences (e.g., biology, physics, and chemistry). Sloman's work on the "space of possible minds" and on architecture-based conceptions of mind is instructive precisely because it treats variation in information-processing architecture as a central dimension along which minded systems may differ (Sloman 1984, 2002).

(A qualification is important here. The difference between narrow and broad pluralism is not simply whether a sequence of mental properties satisfies $\mathcal{P}$. $\mathcal{P}$ is intended to characterize all and only the essential constitutive relations among the standard mental kinds. A sequence might satisfy $\mathcal{P}$ while possessing an essential functional architecture different from that which $\mathcal{P}$ specifies. Its occupants might instantiate some of $\mathcal{P}$'s relations only contingently, so that the same mental properties could exist without satisfying $\mathcal{P}$. Alternatively, they might instantiate additional essential constitutive relations not captured by $\mathcal{P}$. In either case, the sequence satisfies $\mathcal{P}$ without sharing the essential architecture of the standard mental kinds. Such cases belong to broad rather than narrow

pluralism. Narrow pluralism concerns distinct determinate mental sequences sharing the same essential determinable architecture; broad pluralism concerns variation in that architecture itself.)

NP does not, of course, entail BP. There could, in principle, be some deep modal constraints according to which every possible mental sequence, however alien, must instantiate the $\mathcal{P}$ -architecture we instantiate. But NP significantly shifts the intuitive assessment of such a possibility. After all, once it is granted the standard mentals themselves constitute only one kind of mind among others, AU becomes an independent restriction on the space of possible psychology which requires an independent rationale.

Familiar proposals in the philosophy of mind already suggest ways in which psychological organization might depart from the architecture abstracted from the standard mentals. Millikan's (1995) pushmi-pullyu representations combine descriptive and directive functions rather than assigning them to distinct belief-like and desire-like states. Gendler's (2008) aliefs associate representational, affective, and behavioral elements in a state distinguished from belief. Consider also a creature whose doxastic life consists entirely of graded attitudes, with no distinct state occupying the binary-belief role (Buchak 2014; Jackson 2020; cf. Moss 2018). Such a creature might assess evidence, revise its expectations, deliberate, and represent its own doxastic condition, although the relations organizing these activities would differ from those specified by a psychology built around binary belief. Extended-mind cases suggest that information might become available for reasoning and action through pathways involving perception and environmental manipulation where ordinary human cognition proceeds through internal retrieval (Clark and Chalmers 1998).¹²

The significance of these intuitive kinds of cases is that they make alternative forms of psychological organization intelligible. We can consider combining functions, introducing new attitudes, replacing familiar kinds, and reorganizing relations among states without losing our sense that the resulting sequences are psychological. Whether a particular variation preserves the essential architecture specified by $\mathcal{P}$ seems to be inessential to such judgments. There is, moreover, no evident reason why such a revision would eliminate the intensional features that supported attributing mentality to the sequences admitted by narrow pluralism. A credence-only creature could have attitudes whose contents concern those very attitudes. An architecture organized around fused descriptive-directive states could likewise include self-representation and appropriate relations between first-order and higher-order states. Intensional non-well-foundedness does not, by itself, require the specific division of psychological roles found in $\mathcal{P}$.

Shea (2020), responding to Sprevak (2009), defends precisely this kind of move to broad pluralism. He argues that functionalism need not require radically different creatures to instantiate the standard mentals. They may instead possess their own distinct collection of functionally interrelated psychological states. Accordingly, he distinguishes the Ramsey sentences that characterize human psychology ($\mathcal{P}$) from a potentially different set characterizing Martian psychology ($\mathcal{P}_M$).¹³

¹² The point here is not that agent-artifact systems are minded, but merely that our willingness to consider the possibility seriously indicates a pre-theoretical openness to BP.

¹³ Shea frames this separation in terms of the states figuring in each specification: human beliefs are individuated partly by their relations to human desires and intentions (2020, §2). On the Lewisian functionalism he adopts according to which mental terms designate whatever occupies the relevant roles in a given population, this is a

Two differences between the present proposal and Shea's are worth highlighting here. First, Shea's taxonomy is binary. Martians* share our fine-grained functional organization in a different substrate and so possess the standard mentals. Martians** have a wholly different organization and so have psychological states but they are not the standard mentals. On this taxonomy, there is no room here for narrow pluralism's intermediate case, a sequence that instantiates our determinable architecture while consisting of mental kinds distinct from ours. On Shea's picture, occupying the belief role just is believing, so NP collapses into multiple realization. Second, Shea motivates broad pluralism rather than arguing for it. He holds that we should not presuppose that the human inventory exhausts possible psychology, and he points to the octopus as a suggestive actual case. He also explicitly declines to say what it takes to be a psychological state in general. The present argument reaches BP by a different route. Once NP is granted, AU loses its most natural support and stands in need of an independent rationale.

The difference between NP and BP matters most in responding to Sprevak. Sprevak argues that functionalism faces a dilemma. On the one hand, functional roles can be specified more or less finely. If they are specified finely enough to exclude extended processes, they exclude possible Martians whose fine-grained psychology differs from ours. That is, Martians whose memory exhibits no negative transfer, follows different learning curves, or is stored as internal ink marks. On the other hand, if they are specified coarsely enough to admit such Martians, they are also satisfied by agent-artifact systems such as Clark and Chalmers's case of Otto and his notebook. Sprevak adds that the modest extended cognition defended by Clark and Chalmers (1998) cannot be stabilized. By a principle of fair treatment between internal and external cases, every possible internal process licenses its external counterpart. The result is radical extended cognition. By picking up a book I come to implicitly believe everything in it, and a normal subject equipped with a supercomputer has mental arithmetic powers outstripping a calculating prodigy's (2009, 517–18). Sprevak takes these consequences to be absurd and concludes that functionalism is false. Since the pluralism defended here is more liberal about possible minds than the functionalisms Sprevak considers, it may seem particularly susceptible to his argument.

But it is not. The first horn of the dilemma presupposes that admitting alien minds requires coarsening the roles specified by $\mathcal{P}$.¹⁴ NP shows that this presupposition is false along one dimension. *N*-pain and alien-belief are alien in their intrinsic natures, not because the roles they occupy have been loosened. BP, together with Shea's point, shows it is false along the other. Martians who differ from us in fine-grained psychological profile need not satisfy a coarser version of $\mathcal{P}$; they may instead satisfy some distinct $\mathcal{P}_M$. Nothing about such Martians, then, requires that the *human* belief role be specified coarsely, and consequently Otto's notebook fails to satisfy $\mathcal{P}$ in virtue of its distinctive functional position.

Sprevak's argument can, however, be rerun at the level of broad pluralism. Take an agent-artifact system like Otto and imagine a Martian, M, whose internal organization mirrors it. M is minded and so satisfies some $\mathcal{P}_M$.

natural framing (Lewis 1972, 1980). On the non-reductive view assumed here, the Ramsey sentence captures the essential constitutive structure of the mental kinds, and which sequences satisfy it is not settled by the population from whose psychology it was abstracted. This is the same difference that leaves Shea's taxonomy without room for narrow pluralism.

¹⁴ This coarse-grainedness objection is a descendant of Block's (1978) "universal psychology" and the corresponding charge of psychological liberalism.

Generalizing from the argument in section 2, we can assume that every sequence satisfying a genuine psychology is a mental sequence. So if Otto and his notebook satisfy $\mathcal{P}_M$, Otto is alien-minded. To assess this argument, two questions must be separated. The first is which of M's functional relations belong to the psychology in virtue of which M is minded. The second is whether Otto satisfies the psychology so specified.

Sprevak runs these questions together by means of a further principle. If a cognitive agent contains a representation-manipulating process that is significant for guiding its action when employed, that process is one of its cognitive processes (2009, 521; his F2). Given this principle, everything inside M that functions as Otto's notebook functions is thereby psychological.¹⁵ But the principle should be rejected. An organism's total functional organization outruns its psychology. Not every information-bearing process inside a minded creature that contributes to guiding its behavior is thereby a psychological process; the regulatory systems governing thermoregulation or immune response are obvious examples. Nothing about Martians changes this. A minded Martian may contain an ink-mark store that is an internal prop rather than a constituent of its psychology (see fn 18). The genuinely psychological specification of M would then be not the full $\mathcal{P}_M$ but some $\mathcal{P}_M^-$ that omits the store. This verdict respects fair treatment, since the same verdict applies by parity to Otto's notebook.

Sprevak claims that rejecting his principle invites chauvinism, a refusal to count alien processes as cognitive merely because they lack human equivalents (522). But the present criterion is neither location nor resemblance to human cognition. M may have wholly alien processes that are fully mental. What is excluded is whatever does not function psychologically, however human-like it may be. Supplying a general account of what it is for a relation to belong to a psychology is a substantive task, one Shea also leaves open. The present point requires only that the question be distinct, and that Sprevak's principle settles it by *fiat*.

With these two questions separated, Sprevak's construction faces a dilemma of its own. His Martians are otherwise like us, but with an internalized analogue of the external process (519). Either that process belongs to M's psychology or it does not. If it does not, then M's psychology just is a determinate sequence defined by $\mathcal{P}$. M is a mind of standard architecture with an internal visual content, Otto is a mind of standard architecture with an external one, and nothing mental is extended. If it does, further complications arise. Is the added component essential to M's psychology or not? If it is not, it has no place in the psychological specification, which includes all and only the essential constitutive relations. In this case, M's psychology is simply $\mathcal{P}$, and we are returned to the first case. If it is essential, the change does not remain local. In a Ramsified theory, a relation essential to the new component is equally a constraint on the positions to which it relates it, so the occupants of M's belief, desire, and memory

¹⁵ Whether anything inside M functions as Otto's notebook functions is not clear. His ink-mark Martian retrieves information by producing "a mental image of ink marks" in its visual system (508). But imagery and perception are psychologically distinct: imagery is typically under voluntary control and does not present its content as how things are, and so does not ground perceptual judgment as perception does. Otto, by contrast, retrieves information by perceiving his notebook. As described, then, M's retrieval stands in different relations to perception and belief than Otto's, and by the holism of functional specification Otto does not satisfy whatever functional architecture M satisfies. To restore parity, M's retrieval would have to issue in states occupying the perceptual role: internally generated states with the presentational character of visual experience, closer to hallucination, or to an inward-directed sense, than to imagery. But then the store occupies the position of an object of perception in M's psychology rather than a constituent of its memory, and M is simply a mind that perceives an internal prop. Either way, Sprevak's Martian fails to provide the parity his argument requires.

positions must themselves stand in the new relations. By the holism of functional specification, then, M's psychology is not $\mathcal{P}$ with an additional component but a reconfigured $\mathcal{P}_M$ throughout. And Otto, whose internal states occupy the unmodified positions of $\mathcal{P}$, fails to satisfy $\mathcal{P}_M$.¹⁶

Sprevak's dilemma for functionalism therefore dissolves. Alien minds can be taken seriously without coarsening the roles that specify our own, and the inference from possible Martians to actual extended minds fails at each step. Martians who differ from us in fine-grained profile satisfy distinct architectures and the Martians Sprevak constructs to mirror agent-artifact systems either perceive internal props or instantiate psychologies that Otto, whose states are standard mentals, does not satisfy.

Of course, the ability to describe $\mathcal{P}_M$ abstractly does not establish that a natural property sequence could instantiate it. The cases nevertheless provide *prima facie* grounds for regarding the proposed organizations as possible and their occupants as mental. Moreover, AU requires that every such departure from the essential architecture of $\mathcal{P}$ fails as a possible psychology. The familiar proposals countenanced above supply *prima facie* starting points, and the interdependence of functional definitions supplies a mechanism by which they can be understood to be systematic alternatives. Where a proposed revision preserves the grounds for attributing mentality while changing the constitutive organization, there is a positive, if defeasible, reason to regard it as a possible alien psychology.

The apparent conceivability of these alternative architectures suggest that the grounds for attributing mentality can survive changes to the constitutive organization itself. A defense of AU must therefore explain why those changes necessarily destroy mentality despite preserving the general features that supported its attribution in the first place and despite being pitched at the same explanatory and metaphysical level as psychology. Absent such an explanation, AU threatens to reproduce at the level of architecture the chauvinism functionalism was originally designed to avoid. Call this architectural chauvinism. Of course, our own psychology provides a methodological starting point for investigating the space of possible minds. But treating its constitutive architecture as a metaphysical boundary on mindedness is a much less defensible position.¹⁷

4. Philosophical implications

The possibility of both forms of alien minds immediately raises a host of thorny conceptual issues in the philosophy of mind which have significant knock-on impacts across philosophy more generally. We can begin to see these

¹⁶ This leaves open the possibility that Otto might instantiate a nonstandard mental kind whose own nature involves relations to the standard mentals, though theirs involves no relation to it, as Shea (2020, §3) suggests when he raises the possibility of a category of readily-accessible artifact-based memory. Whether the notebook state is an instance of such a kind turns on whether it genuinely functions psychologically, a question the present argument need not settle. So too does the question whether such states would belong to Otto or to the hybrid system (cf. Miyazono 2017). But even if there are such states, they are not beliefs, and none of Sprevak's absurdities, all formulated in the vocabulary of the standard mentals, follow.

¹⁷ The distinction between narrow and broad pluralism also helps locate different questions within the wider literature on alien minds. These include the limits of projecting human psychological categories onto animals (de Waal 1999; Andrews 2012, 2020), the possibility of unfamiliar psychological kinds in extraterrestrial or artificial minds (Schneider 2015, 2019), and variation in cognitive architecture within the space of possible minds (Sloman 1984, 2002).

impacts by considering the departure point of this paper—the non-reductive functional definitions of determinable mental properties under which determinate mental properties (including the standard mentals) fall. Even though the general metaphysical structure is familiar, these properties already involve a bit of conceptual engineering, broadly understood. The failure of Bealer’s non-reductive definitions seemed to require “building out” new conceptual space to accommodate the class of possible minds contemplated by NP since our ordinary conception of the mental properties doesn’t clearly distinguish (if it distinguishes it at all) a level of mental determinables of which the standard mentals are but one determinate. NP, in turn, raised in a particularly acute way the possibility of architectural chauvinism (and, hence, BP). And while imagining alien cognition is not foreign even in pop culture, it isn’t clear how exactly it is to be understood.¹⁸

These developments, however, merely set the stage for the revisionary work in store. Narrow pluralism doesn’t just give us the possibility of nonstandard minds, but a range of nonstandard mental properties that differ intrinsically from the standard mentals, including with respect to phenomenology or the lack thereof. Similarly, broad pluralism gives us an open-ended range of nonstandard mental properties and relations that differ functionally from the standard mentals.¹⁹ The possibility of these kinds of alternative possible minds has implications that propagate beyond the philosophy of mind itself, at least insofar as many of our philosophical theories purport to hold generally. In this concluding section, I illustrate with two cases. The first, from epistemology, turns on narrow pluralism (§4.1). The second, from the theory of agency, turns on broad pluralism (§4.2). The intention in these remarks is not to decisively argue for any particular view, but rather to illustrate by way of example how acceptance of both kinds of mental pluralism affects philosophical theorizing.

4.1. Phenomenalism and narrow mental pluralism

A significant amount of recent work in epistemology seems to make essential reference to phenomenological properties in order to account for the justification of our beliefs. For specificity, let’s consider moderate dogmatism (e.g., Pryor 2000; Huemer 2001) as a representative example:

MD: If x has a seeming as of p ’s being the case, then x thereby has some degree of immediate (*prima facie*) justification for believing that p .²⁰

On one standard way of developing the view, MD holds in virtue of the *presentational phenomenology* of seemings (see Bengson 2015, 2023, Chudnoff 2013, and Koksvik 2021 among others). Call this further thesis “Phenomenalism”.

We argued in §2 that there appear to be possible, fully minded beings who lack phenomenology entirely, both zombies and (potentially) Eklund-style aliens. Call these “aliens_{np}”. Their internal states satisfy $\mathcal{P}$ and so count

¹⁸ Consider, for instance, the character Data (from *Star Trek: The Next Generation*). These characters are portrayed simply as having some range of the standard mental states associated with intelligence, but lacking many of the standard emotive states. As we have seen, however, the mental holism intrinsic to functional definitions makes this pretense unworkable.

¹⁹ Understanding the mental properties within this conceptual space even in general terms involves us in some of the most difficult problems in cognitive ethology without the benefit of phylogenetic or ontogenetic constraints to guide us (Allen & Bekoff 1997, Bekoff, Allen, & Burghardt 2002).

²⁰ Essentially the same points can be made regarding, for example, Smithies’ access mentalism (2012, 2019).

as genuinely mental. If $\text{aliens}_{\text{np}}$ can be justified, Phenomenalism faces some significant philosophical challenges, which I will elaborate below. But first, let's consider the natural response, namely, denying that they can be justified. This is, I suspect, the route most phenomenologists will initially reach for and, indeed, may think that they have already defended.

They have not. The reason is given by the argument in §2 for coordinated substitution into the mental sequence. Since the narrow pluralist is not committed to holding any of the standard mental determinates fixed in the determinate sequence, any argument that runs from the necessity of the standard mental determinates requiring phenomenology for justification to the conclusion that *any* sequence of mental determinates satisfying $\mathcal{P}$ requires phenomenology for justification is inadequate. This, in turn, will likely require an argument that phenomenology is the only possible mechanism for the kind of access to our mental lives required for justification. To my knowledge, this very general argument has never been attempted.²¹

And there are reasons to be skeptical of such a generalization. The general difficulty with this view can be put as a dilemma. Either phenomenology makes a difference to the characteristic psychological relations through which a subject accesses and responds to its own cognitive states, or it does not. If it does, the difference belongs to the functional organization captured by $\mathcal{P}$. But $\text{aliens}_{\text{np}}$ instantiate that organization through distinct, non-phenomenal mental properties. Their introspective capacities can access their $\text{seemings}_{\text{np}}$, and they can recognize and respond to those states as reasons for $\text{belief}_{\text{np}}$. Phenomenology is therefore not the only possible basis for the relevant epistemic organization.

Suppose, alternatively, that phenomenology makes no difference to these relations. The phenomenologist may then maintain that phenomenology nevertheless provides a distinctive, noncausal normative ground for justification. But this response sits uneasily with the internalist motivations for Phenomenalism. The familiar appeal to presentational phenomenology is motivated by the thought that seemings make their contents available to the subject as reasons for belief, rather than merely causing belief. Yet the alien_{np} possesses the corresponding psychological organization in that its introspective capacities provide $\text{access}_{\text{np}}$ to its own states, and it can form higher-order judgments about their evidential significance, respond to apparent defeaters, and revise its $\text{beliefs}_{\text{np}}$ accordingly. The phenomenologist may insist that presentational phenomenology supplies a distinctive normative ground absent from the alien's psychology. But once the psychological relations involved in $\text{accessing}_{\text{np}}$, $\text{assessing}_{\text{np}}$, and responding to one's own states are held fixed, it is no longer clear why that additional phenomenal feature is indispensable to justification. Explaining its indispensability requires more than appealing to the access-and-guidance capacities that originally motivated Phenomenalism, since those capacities appear to survive coordinated substitution.

Either phenomenology makes a distinctive difference to the cognitive roles its bearers play, or it does not. If it does, the difference will show up in the pattern of relations isolated by Ramsification, and since the states of $\text{aliens}_{\text{np}}$ satisfy $\mathcal{P}$, alien_{np} cognition preserves it. In which case, we have reason to infer that phenomenology is not the only possible mechanism. If it does not, phenomenology contributes to justification, if at all, as some form of

²¹ See, for example, Smithies (2019) on access, Horgan and Tienson (2002), Kriegel (2011), and Mendelovici (2018) on content, and Bengson (2015, 2023) and Chudnoff (2013) on presentationality.

non-causal ground. It makes no difference to how the subject's states are accessed, how they guide belief, or how they respond to evidence. In which case, the phenomenology itself is idle even in our own case.

The dilemma can be restated as a choice about what justification supervenes on. Because *aliens_{np} satisfy $\mathcal{P}$, their psychological states instantiate the same determinable functional architecture as ours, although the determinate properties and relations occupying its roles differ*. A phenomenalist who denies them justification must therefore either deny that *aliens_{np} satisfy $\mathcal{P}$* (contrary to §2) or hold that justification supervenes not on that determinable architecture alone but on the architecture together with phenomenology. Some internalists may accept the second option, since the supervenience base remains wholly mental. But the usual motivations for Phenomenalism are access and guidance—justification must be available to the subject and capable of guiding belief from the subject's own point of view. On the second option, however, phenomenology makes no difference to the determinable relations through which a subject accesses, assesses, and responds to its own psychological states. An *alien_{np}* and its phenomenal counterpart instantiate corresponding forms of introspection, higher-order judgment, and belief revision, including judgments concerning the apparent evidential significance of their own states. Yet the phenomenalist must hold that one is justified and the other is not. The difference in epistemic standing therefore turns on features of the determinate psychological properties that make no difference to the determinable organization of those capacities. Why those features should be decisive for justification, when the internalist motivation for Phenomenalism appeals to access and guidance, requires further explanation.

One response to this problem is to go parochial about justification and restrict it to creatures with phenomenological access. Even if this restriction can be motivated, it raises the question: Can *aliens_{np}* be epistemically evaluated in any sense? It is one thing to claim that they necessarily lack a specific epistemic property, justification, that only phenomenally endowed beings can have. It is another to claim that they lack any epistemic standing whatsoever. Once *$\mathcal{P}$ -satisfaction* is granted, the second claim is very hard to sustain. *Aliens_{np} form beliefs_{np} in response to evidence, draw inferences_{np}, and revise their views_{np} in light of counterevidence*. Given the self-consciousness relations in *$\mathcal{P}$* , they can also ask themselves whether they ought to believe_{np} what they do. Their cognition can go well or badly in the same ways ours can. An *alien_{np}* that ignores a defeater or generalizes hastily does exactly what we would criticize without hesitation in one another. To deny that any of this is an epistemic failure, stretches credulity. Rather, a phenomenalist who denies *aliens_{np} justification* should grant that they have some positive epistemic standing. The question is what that standing is.

There are two possibilities. First, *aliens_{np}* might have a standing distinct from justification but on a par with it; roughly, a property that occupies the same normative role (in the present example: immediate, *prima facie* standing to believe_{np}) on a different ground. This yields a form of modal disjunctivism according to which the ground of immediate justification varies across possible minds. But it is unclear how stable this view is. What makes the *alien_{np}* standing a counterpart of justification? Its normative profile is fixed by the cognitive role of the states that bear it. The *seemings_{np}* of *aliens_{np}* occupy the same cognitive role as ours: they present_{np} their contents, incline_{np} the subject to believe_{np}, and are defeated by counterevidence in the same way as *seemings* behave in us. So, the standing they confer licenses the corresponding cognitive states in the same ways that justification does. The two properties

share a normative profile because their bearers share a cognitive role. That is, the modal disjunctivist must nevertheless deny that these very close commonalities ground a single epistemic property.

The second possibility is that $\text{aliens}_{\text{np}}$ have an epistemic standing we share with them. Their $\text{seemings}_{\text{np}}$ would then confer that standing in virtue of something other than phenomenology. This preserves a kind of epistemic unity between us and them. But this option is also unstable. Once we grant that we share with $\text{aliens}_{\text{np}}$ a normative property that their seemings confer without phenomenology, we are forced to consider the ground of this property. Whatever the answer (the cognitive role specified by $\mathcal{P}$, the reliability of the states that occupy it, or something else), our seemings have that feature as well. Suppose, for instance, that the standing of $\text{aliens}_{\text{np}}$ is reliabilist (Levy 2024). It is then unclear why reliability is not what does the basic epistemic work in our case too, with phenomenology an accompaniment rather than a ground. The phenomenalist can retreat to the claim that justification is a further, phenomenally grounded standing layered on top of the shared one. But this reproduces the dilemma above. This shared epistemic standing already fixes what the subject's seemings license, what defeats them, and how the resulting beliefs are assessed, so the further epistemic features appear to make no difference (Current Author; cf. Butzer 2026).

The point here is not to attempt a decisive refutation of Phenomenalism. Rather, it is illustrative. Once we admit complex alien minds that differ significantly from our own, the project of epistemology becomes increasingly difficult and abstract. Moreover, it seems unlikely that at this level of abstraction our ordinary epistemic concepts will be fully adequate to characterize the field. This points to a need for conceptual engineering guided by what I call “conceptual productivity” [Current author]. Understanding a subject involves grasping facts and their interrelations (Bengson 2017), but even a partial understanding can organize a domain in ways that place rational pressure on an agent to extend their conceptual repertoire. A conception of a given subject is conceptually productive when more fully understanding the subject requires grasping and applying a novel concept, without which the relations among its central features cannot be adequately captured. In the present context, reflection on alien minds may require more than merely extending familiar epistemic concepts to unfamiliar cases. It may require developing new concepts necessary for us to adequately understand epistemology itself.

4.2. Agency and broad mental pluralism

The problems discussed in the preceding section are not simply an artifact of a well-chosen case. They represent problems of generalization that are likely to arise in a wide range of philosophical contexts. In this section, I briefly consider a similar set of issues in the philosophy of agency.

Narrow pluralism puts pressure on theories that build in the intrinsic character of the standard mentals. Broad pluralism puts pressure on theories that build in their architecture, and the theory of agency offers a clear case study. On the standard causal theory, an action is behavior caused, in the right way, by a belief–desire pair that rationalizes it (Davidson 1963). On the Humean theory of motivation, being motivated requires a desire together with a means–end belief (Smith 1994). Both analyze agency in terms of a division of labor between states with opposite directions of fit. That division is a feature of our architecture, and broad pluralism allows for minds that lack it. Millikan (1995) argues that the most basic representations are “pushmi-pullyu” representations, single states that at once describe how things are and direct what is to be done. Her examples are simple, but they show that the

division of direction of fit is not a precondition of representation. A sophisticated architecture could be built from such states, with whatever further integration and self-monitoring mentality requires, without ever separating belief from desire the way the standard mentals do.

Could such beings act? Their behavior can be flexible, goal-directed, and responsive to changing circumstances, and by hypothesis it issues from genuinely mental states. Yet they fail the belief–desire analysis, not because their states are poorly organized but because they lack the states the analysis requires. One might respond by individuating belief and desire more coarsely, so that any state with a world-directed aspect counts as a belief and any state with a goal-directed aspect counts as a desire. This faces two problems. First, roles coarse enough to include such beings threaten to include far too much, thus facing the problem of liberalism that has long dogged functionalism (Block 1978; cf. Sprevak 2009). The second is specific to this case. Even coarse belief and desire roles are two roles, and the analysis requires two states, one occupying each. A pushmi-pullyu architecture supplies only one. Counting a single state as both a belief and a desire preserves the letter of the analysis while abandoning its explanatory content, which lies in the interaction of states with distinct directions of fit. A defender who coarsens this far has, in effect, already given up the belief–desire theory for something like the determinable account of §3.

Given this, the options discussed in §4.1 recur. (A) One can deny that such beings act, but that verdict needs independent motivation if it is not simply to relocate the parochialism. (B) One can treat agency as disjunctive, grounded in belief–desire pairs in us and in something else in them. This move inherits the instability of disjunctivism noted above. Or (C), one can look for a more general property: roughly, behavior guided by states that represent both how things are and what is to be done, whether or not those representations divide into distinct states. For example, a defender of the belief–desire theory might propose that what matters to agency is the distinction between descriptive and directive functions, not their realization in distinct psychological states. This is a reasonable generalization, but it concedes the present point: the resulting analysis no longer presupposes the functional architecture of the standard mentals. On this approach, the belief–desire theory is not refuted but redescribed as an account of one way of realizing agency.

Conclusion

Philosophy, in general, and philosophy of mind, in particular, begin with the standard mentals because they are the psychological properties with which we are familiar. But their methodological priority provides no reason to regard them as exhausting the possibilities for mentality. I have argued for two forms of mental pluralism. Narrow pluralism holds that the functional architecture of the standard mentals may be instantiated by distinct sequences of mental properties, differing from ours in their intrinsic natures. Broad pluralism goes further in allowing possible minds whose essential functional organization differs from our own. Together, these arguments challenge the assumption that the psychological properties and relations familiar from human cognition fix the boundaries of mentality.

The consequences of pluralism, however, extend well beyond the philosophy of mind. Narrow pluralism raises difficulties for philosophical theories that depend on the intrinsic character of the standard mentals (as illustrated by phenomenological accounts of epistemic justification). Broad pluralism raises a corresponding problem for theories that presuppose their characteristic functional organization (as illustrated by belief–desire

accounts of agency). Neither difficulty necessarily refutes the theories in question. Rather, they raise questions about their scope and about the grounds on which they can be generalized to unfamiliar kinds of minds.

The resulting methodological problems are, nevertheless, substantial. The central challenge is not simply in extending familiar categories to unfamiliar cases but determining whether those categories capture the appropriate distinctions in the first place. The conceptual joints visible in human psychology need not be the joints at which a general philosophical theory should be carved.

From a purely theoretical perspective, these difficulties might seem tolerable. Even a fully adequate theory of human justification, agency, or welfare would be a substantial achievement, and the problem of generalizing beyond the human case might be regarded as a downstream aspiration. But the development of increasingly sophisticated artificial systems has made these questions practical as well as theoretical. There is no guarantee that artificial minds, should they emerge, will instantiate either the standard mental or their familiar functional architecture.²² The challenge of understanding and evaluating such systems therefore extends beyond determining whether familiar psychological and normative properties apply to them. It may require identifying alternative forms of those properties, or developing more general categories in which both human and artificial cases can be understood. These questions have immediate implications for AI alignment (Josifović & Noller 2026; Current Author) and AI welfare (Shepherd 2024; Goldstein & Kirk-Giannini 2025; Veit 2026).

It is perhaps darkly ironic that philosophical speculation about zombies and alien minds, long regarded as paradigmatically impractical speculation, has become central to some of the most pressing practical questions of our age.

References

- Allen, C., & Bekoff, M. (1997). *Species of Mind: The Philosophy and Biology of Cognitive Ethology*. MIT Press.
- Andrews, K. (2012). Beyond anthropomorphism: Attributing psychological properties to animals. In T. L. Beauchamp & R. G. Frey (eds.), *The Oxford Handbook of Animal Ethics* (pp. 469–494). Oxford University Press.
- Andrews, K. (2020). *How to Study Animal Minds*. Cambridge University Press.
- Båve, A. (2017). Self-consciousness and reductive functionalism. *The Philosophical Quarterly*, 67(266), 1–21.
- Bealer, G. (1997). Self-consciousness. *The Philosophical Review*, 106(1), 69–117.
- Bealer, G. (2000). Fregean equivocation and Ramsification on sparse theories: Response to McCullagh. *Mind & Language*, 15(5), 500–510.
- Bealer, G. (2001). The self-consciousness argument: Why Tooley's criticisms fail. *Philosophical Studies*, 105(3), 281–307.
- Bealer, G. (2010). The self-consciousness argument: Functionalism and the corruption of content. In R. C. Koons & G. Bealer (eds.), *The Waning of Materialism* (pp. 137–158). Oxford University Press.
- Bekoff, M., Allen, C., & Burghardt, G. M. (eds.). (2002). *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition*. MIT Press.
- Bengson, J. (2015). The intellectual given. *Mind*, 124(495), 707–760.
- Bengson, J. 2017. The unity of understanding. In Grimm (Ed.), *Making Sense of the World*. New York: Oxford.

²² For recent discussions, in addition to Sloman (1984, 2002), see Cappelen & Dever (2024), Lewis & Sarkadi (2024), and Wong (2025). Schneider (2015, 2019) makes the point more generally.

- Bengson, J. (2023). Intuition in philosophical inquiry. In McCain, Stapleford, & Steup (eds.), *Seemings*. New York: Routledge.
- Block, N. (1978). Troubles with functionalism. In C. W. Savage (ed.), *Perception and Cognition: Issues in the Foundations of Psychology* (Minnesota Studies in the Philosophy of Science, Vol. 9, pp. 261–325). University of Minnesota Press.
- Buchak, L. (2014). Belief, credence, and norms. *Philosophical Studies*, 169(2), 285–311.
- Butzer, T. (2026). On the epistemic insignificance of perceptual phenomenology. *Analytic Philosophy*, 1–12.
- Cappelen, H., & Dever, J. (2024). AI with alien content and alien metasemantics. In E. Lepore & L. Anderson (eds.), *The Oxford Handbook of Applied Philosophy of Language* (pp. 573–593). Oxford University Press.
- Chalmers, D. 1996, *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford.
- Chudnoff, E. (2013). *Intuition*. Oxford: Oxford University Press.
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19.
- Davidson, D. (1963). Actions, reasons, and causes. *The journal of philosophy*, 60(23), 685–700.
- de Waal, F. B. M. (1999). Anthropomorphism and anthropodenial: Consistency in our thinking about humans and other animals. *Philosophical Topics*, 27(1), 255–280.
- Eklund, M. (2026). Alien minds: The case of consciousness. Paper presented at the After ‘Consciousness’ Workshop, University of Hong Kong, January 27–28, 2026. <https://ai-humanity.net/alien-minds-the-case-of-consciousness/>.
- Funkhouser, E. 2023. Determinate/determinable. In Fisher & Maurin, *The Routledge Handbook of Properties*. New York: Routledge.
- Gendler, T. S. (2008). Alief and belief. *The Journal of Philosophy*, 105(10), 634–663.
- Goldstein, S., & Kirk-Giannini, C. D. (2025). AI wellbeing. *Asian Journal of Philosophy*, 4, Article 25.
- Horgan, T., and J. Tienson (2002). The intentionality of phenomenology and the phenomenology of intentionality. In D. J. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*. New York: Oxford University Press.
- Huemer, M. (2001). *Skepticism and the Veil of Perception*. Bloomsbury Publishing PLC.
- Jackson, E. G. (2020). The relationship between belief and credence. *Philosophy Compass*, 15(6), e12668.
- Josifović, S., & Noller, J. (2026). Agency and alignment: Toward a normative architecture for human–AI interaction. *AI & Society*, 41, 5561–5571.
- Koksvik, O. (2020). *Intuition as conscious experience*. New York: Routledge.
- Kriegel, U. (2011). *The Sources of Intentionality*. Oxford: Oxford University Press.
- Lee, G. (2019). Alien subjectivity and the importance of consciousness. In A. Pautz & D. Stoljar (eds.), *Blockheads! Essays on Ned Block’s Philosophy of Mind and Consciousness* (pp. 215–242). MIT Press.
- Levy, N. (2024). Consciousness ain’t all that. *Neuroethics*, 17(2), Article 21.
- Lewis, D. (1970). How to define theoretical terms. *The Journal of Philosophy*, 67(13), 427–446.
- Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 50(3), 249–258.
- Lewis, D. (1980). Mad pain and Martian pain. In Block (Ed.), *Readings in the Philosophy of Psychology* (Vol. 1). Cambridge, MA: Harvard University Press.
- Lewis, P. R., & Sarkadi, Ş. (2024). Reflective artificial intelligence. *Minds and Machines*, 34, Article 14.
- McCullagh, M. (2000). Functionalism and self-consciousness. *Mind & Language*, 15(5), 481–499.
- Mendelovici, A. A. (2018). *The Phenomenal Basis of Intentionality*. Oxford: Oxford University Press.
- Millikan, R. G. (1995). Pushmi-pullyu representations. *Philosophical Perspectives*, 9, 185–200.
- Miyazono, K. (2017). Does functionalism entail extended mind?. *Synthese*, 194(9), 3523–3541.
- Moss, S. (2018). *Probabilistic Knowledge*. Oxford University Press.

- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450.
- Pryor, J. (2000). The skeptic and the dogmatist. *Noûs*, 34(4), 517–549.
- Schneider, S. (2015). Alien minds. In S. J. Dick (ed.), *The Impact of Discovering Life beyond Earth* (pp. 189–206). Cambridge University Press.
- Schneider, S. (2019). *Artificial You: AI and the Future of Your Mind*. Princeton University Press.
- Shea, N. (2020). Functionalism interrelations among human psychological states inter se, ditto for Martians. In J. Smortchkova, K. Dołęga, & T. Schlicht (eds.), *What Are Mental Representations?* (pp. 242–253). Oxford University Press.
- Shepherd, J. (2024). Sentience, Vulcans, and zombies: The value of phenomenal consciousness. *AI & Society*, 39(6), 3005–3015.
- Shoemaker, S. 1982. The inverted spectrum. *The Journal of Philosophy*, 79(7), 357–381.
- Sloman, A. (1984). The structure of the space of possible minds. In S. Torrance (ed.), *The Mind and the Machine: Philosophical Aspects of Artificial Intelligence*. Ellis Horwood.
- Sloman, A. (2002). Architecture-based conceptions of mind. In P. Gärdenfors, K. Kijania-Placek, & J. Woleński (eds.), *In the Scope of Logic, Methodology and Philosophy of Science, Vol. II*. Kluwer Academic Publishers.
- Smith, M. (1995). Internal reasons. *Philosophy and phenomenological research*, 55(1), 109–131.
- Smithies, D. (2012). The mental lives of zombies. *Philosophical Perspectives*, 26(1), 343–372.
- Smithies, D. (2019). *The Epistemic Role of Consciousness*. Oxford University Press.
- Sprevak, M. (2009). Extended cognition and functionalism. *The Journal of Philosophy*, 106(9), 503–527.
- Tooley, M. (2001). Functional concepts, referentially opaque contexts, causal relations, and the definition of theoretical terms. *Philosophical Studies*, 105(3), 251–279.
- Veit, W. (2026). Is consciousness required for AI welfare? *Asian Journal of Philosophy*, 5, Article 18.
- Wong, H. Y. (2025). Interrogating artificial agency. *Frontiers in Psychology*, 15, 1449320.
- Yablo, S. 1992. Mental causation. *The Philosophical Review*, 101, 245–280