

Taking Alien Minds Seriously

Nonstandard Cognition & the Scope of Philosophy

Marc A. Moffett

Draft 9/29/2026 — Do not cite without permission

Abstract

Philosophy of mind begins from familiar human psychological kinds: belief, desire, perception, intention, pain, and their kin. This paper argues that their methodological priority should not be mistaken for a metaphysical constraint on possible minds. I begin with Bealer's Self-Consciousness Argument. Once reductive functionalism is abandoned and mental properties themselves are allowed to occupy functional roles, non-reductive functional definitions fail to uniquely isolate the standard mentals: distinct natural mental sequences may instantiate the same determinable architecture while differing in their intrinsic natures. This motivates narrow mental pluralism. I then argue that narrow pluralism undermines the principal motivation for architectural uniqueness. If the standard mentals do not exhaust mentality, a fixed functional architecture abstracted from them cannot simply be assumed to exhaust possible psychology. This motivates broad mental pluralism and a corresponding objection to architectural chauvinism. Finally, I trace the implications beyond philosophy of mind. Theories of justification, agency, welfare, and related properties often aspire to generality while relying on features of the standard mentals. Alien minds therefore generate a conceptual-engineering problem: philosophy must identify the level of abstraction at which its categories genuinely generalize.

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

Descartes asked what he could know with certainty. His answer: "I think." Whether or not one accepts Descartes's conclusion, there is little doubt that we know much about our mental lives with a directness and certitude that much of our understanding lacks. And yet, this very familiarity may be an important source of philosophical error. Indeed, it is not implausible to say that the philosophy of mind has always faced a problem of parochialism. The minds with which we are most intimately acquainted are human minds, and the psychological properties about which we theorize are correspondingly the familiar ones: belief, desire, perception, memory, intention, pain, emotion, imagination, and their kin. Call these *the standard mentals*.

The philosophy of mind has been largely concerned with the standard mentals. This is unsurprising. It is perhaps the only feasible starting point for inquiry into the mind. But that is a methodological fact, not a metaphysical one. The methodological priority of the home case is unobjectionable, but it is not without its philosophical dangers. Focusing on the human case can obscure not just issues in the philosophy of mind itself, but also issues across a broad range of topics in ethics, action theory, and epistemology that intersect with it.

This paper develops the case for two forms of mental pluralism and then considers their broader philosophical consequences. The argument begins with Bealer's Self-Consciousness Argument (1997, 2001, 2010). Reductive functionalism was aimed at a specific philosophical project, namely, providing a physicalistically acceptable definition of the standard mentals consistent with multiple realizability. Bealer's argument shows that the reductive side of this project fails. Despite this, Bealer attempted to preserve its definitional side by jettisoning the restriction of the quantifiers to first-order physical properties and allowing the mental properties themselves to occupy the functional roles. If the standard mentals uniquely satisfy the resulting Ramsey sentences, the result will be non-reductive functional definitions of the mentals. Evidently, however, this uniqueness claim fails. Distinct sequences of natural properties can instantiate the same psychological organization while differing in their intrinsic natures. I argue that this gives us positive reason to accept the possibility of mental properties beyond the standard mentals and motivates what I call *narrow mental pluralism*.

The argument then considers what follows once narrow pluralism is granted. Narrow pluralism remains conservative in one important respect: every admissible mental sequence must instantiate the determinable functional architecture abstracted from the standard mentals. But the most obvious rationale for that restriction was precisely the thought that the standard mentals exhaust mentality. Once that thought is abandoned, the architecture abstracted from those kinds loses its automatic claim to exhaust psychology. Moreover, modest reorganizations of the relevant functional definitions appear capable of generating non-isomorphic "psychologies" whose states have the same sort of metaphysical and explanatory credentials as the alien states already admitted by narrow pluralism. Narrow pluralism does not logically entail this kind of *broad mental pluralism*. It does, however, shift the burden of proof. Anyone who accepts alien mental occupants while insisting that every possible mind must preserve the human-derived architecture owes an account of why architectural variation, unlike variation in occupants, destroys mentality.

The final section turns to the broader philosophical consequences. If narrow and broad pluralism are live possibilities, then philosophical theories that appeal to the standard mentals face a problem of scope: even theories adequate for human justification, agency, or welfare may fail to generalize to nonstandard minds. Addressing that problem requires conceptual engineering at two levels: (1) articulating the relevant space of possible minds well enough to determine whether home theories generalize; and (2) where they do not, determining whether the higher-level epistemic, agential, or normative categories themselves need to be reanalyzed (or regrounded) or replaced by more general ones. The paper is therefore modal rather than predictive. I do not argue that such alien minds are actual, but that their possibility should be treated as a working hypothesis rather than ruled out by the structure of our own psychology.

1. Reductive functionalism and the Self-Consciousness Argument

Standard functionalism seems, at first blush, to offer the natural response to this parochialism. The great attraction of functionalism is that it separates mentality from any particular physical substrate. Martian hydraulics, silicon circuitry, and biological neurons may all realize the same mental state if they occupy the appropriate place in the relevant causal or psychological organization. But the familiar multiple-realizability lesson concerns variation beneath the psychological taxonomy. It tells us that pain need not be realized in carbon. It does not yet tell us

whether every possible psychology must contain pain, belief, desire, perception, and the rest—or even determinable states abstracted from those familiar kinds.

On the standard functionalist recipe, we begin with a psychological theory P whose theoretical vocabulary includes predicates for the standard mentals. P may be assembled from the platitudes of folk psychology, as in analytic functionalism, or from an empirical psychological theory, as in psychofunctionalism (Lewis 1970, 1972; Block 1978). Ramsification then replaces the mental predicates, or nominalized forms of them, with variables that are existentially quantified over. This allows us to isolate the characteristic pattern of causal and psychological relations of the standard mentals abstracted from P . On reductive functionalism, the quantifiers are restricted to first-order, typically physical, realizer states and the mental properties are thereby defined as higher-order properties.

Bealer’s Self-Consciousness Argument (SCA) targets the fact that psychological theory is *intensionally non-well-founded* (Bealer 1997, 2001, 2010). A canonical instance is a self-intimation principle of roughly the following form:

$\mathcal{J}$. 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 can occur in an intensional position as part of the content of that very property. This creates a dilemma for the reductive functionalist. To see why, consider two options for generating the Ramsey sentence for $\mathcal{J}$. On the standard functionalist recipe, we remove all of the theoretical (in this case, psychological) vocabulary from $\mathcal{J}$, replace it with appropriate variables, and bind those variables with quantifiers restricted to first-order physical realizer states:¹

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

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 can play the “belief role”. 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 even if there were such a state, 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:²

¹ For simplicity, ignoring the rest of $\mathcal{P}$. Bealer extends the argument to Lewis’s realizer 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 The example isn’t intended as a genuine candidate for including in the Ramsey definitions.

$\mathcal{J}$. x believes that pain is a psychological state.

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

$J_R. (\exists r_1, r_2)(x \text{ } r_1 \text{ } s \text{ that } r_2 \text{ is a psychological state})$

Quantifying into the embedded clause distorts the pattern of the mentals that the functional definitions are supposed to isolate. The problem is that 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.³

What goes for $\mathcal{J}$, however, 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 neurological state of human beings. 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 accidental. Psychological theories evidently can contain attitudes about 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.

The SCA does not show that functional organization is irrelevant to mentality. It shows that the standard reductive strategy cannot define the standard mentals by replacing them with ontologically prior realizers while preserving the psychological organization from which the definitions were abstracted. 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 is not obviously fatal to functionalism proper. Bealer’s response was to go non-reductive. In non-reductive functional definitions, we remove the restriction to ontologically prior first-order realizers and allow the quantifiers to range over the mental properties themselves (Bealer 2010, 154–58). Since the definitions are abstracted from the standard mentals, they trivially satisfy them. Bealer notes that if no other natural, non-Cambridge properties satisfy the resulting definitions, then the definitions uniquely isolate the standard mentals. 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

³ 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.

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. But this proposal about the psychological level should be distinguished from the Uniqueness Hypothesis. It is one thing to hold that primitive intensional non-well-foundedness is characteristic of the psychological level; it is another to hold that the standard mentals are the *only* natural mental properties capable of exhibiting it. The latter claim is considerably stronger and, I will argue, problematic.

To begin with, consider a creature who has a state with the characteristic causes and effects associated with our pain but whose phenomenal character is nausea-like. Or consider a creature whose corresponding state has a phenomenal character entirely alien to us. Finally, we can even consider a creature whose pain-role state lacks phenomenal character altogether.⁴ Such property sequences appear at least *prima facie* possible and do not seem to be excluded merely by restricting the range of the variables to natural properties. Nor do the cases seem to undermine the proposal that the relevant occupants are mental. Instead, they undermine specifically the Uniqueness Hypothesis; that is, the claim that satisfying the functional pattern uniquely identifies the standard mentals that occupy those roles in us.

Assuming that at least some of the standard mentals (e.g., pain) have their phenomenal natures essentially (cf. Smithies 2019), this point can be made especially clearly. If, say, pain has its phenomenal character essentially, then a state like the first one hypothesized above (call it *n*-pain) that differs from pain only in phenomenal character is not, itself, pain. So, functional equivalence does not entail property identity; sameness of role can't serve to individuate mental properties. A non-reductive functional characterization of pain therefore fails to isolate pain uniquely and so fails to provide an adequate definition of it.

The point generalizes. A sequence of mental properties could instantiate the pattern determined by psychological theory $\mathcal{P}$ while differing, property by property, from the standard mentals that instantiate that pattern in us. The very move that preserves functionalism in the face of the SCA, therefore, creates trouble for the Uniqueness Hypothesis. Once the occupants of the functional roles are allowed to be irreducibly mental, we can no longer infer simply from occupancy of the belief-role, pain-role, desire-role, or perception-role that the occupant is belief, pain, desire, or perception.

The spirit of Bealer's non-reductive proposal can nevertheless be preserved by introducing a determinable/determinate structure (Moffett 2025). Rather than treating Ramsification as defining the standard mentals, we can instead use the characteristic pattern exhibited by them to characterize *determinable* functional-mental properties under which those states fall. Other determinate mental sequences may instantiate the same characteristic pattern while differing from the standard mentals.

⁴ This possibility does not commit us to phenomenological zombies tout court, though I think that possibility does ultimately follow. For now, it merely admits of non-phenomenological realizers of the pain role itself. I address this case in more detail below.

The determinate properties occupying the relevant positions must instantiate the psychological pattern in the ordinary functionalist sense. Thus, each admissible determinate sequence must satisfy principles such as *J* and consequently exhibit the characteristic intensional non-well-foundedness we have been considering. The determinable properties abstracted from those sequences, however, do not and need not themselves satisfy those principles. One does not, for example, determinably believe that one determinably believes that *p*. The determinables characterize a common functional architecture that admissible determinate sequences instantiate. The occupants are the irreducible determinate mental properties themselves.

The standard mental states constitute one such determinate sequence. Individual determinates can then be characterized partly by the distinctive way in which they occupy the relevant determinable functional positions. In the pain case, for example, phenomenal character may partly explain the way in which the determinate state pain occupies the determinable pain-role. This does not imply the phenomenal character itself is sufficient for occupying the role. A state might have the phenomenal character normally associated with pain without occupying the pain-role and hence without being pain. Conversely, a state like *n*-pain with the phenomenal character normally associated with nausea might nevertheless occupy the determinable pain-role, perhaps partly in virtue of that character.

Assuming that pain and *n*-pain have their respective phenomenal characters essentially, they are distinct mental states; that is, distinct mental determinates falling under the same determinable functional-mental property. So, the pain/*n*-pain case implies what we might call Weak Narrow Mental Pluralism:

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

In the present case, the new sequence concerns a single property substitution. Moreover, *n*-pain has, by hypothesis, a familiar phenomenal character. Consequently, it is difficult to see how one could plausibly deny that a sequence containing *n*-pain in the pain-position is itself a sequence of mental properties.

The more interesting question is whether the same conclusion extends to every natural determinate sequence that instantiates the relevant psychological architecture. Call the stronger thesis General Narrow Mental Pluralism, or simply Narrow Pluralism:

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

Two considerations support the generalization from WNMP to NP. First, every admissible determinate 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 *P*, where *P* describes the essential functional architecture of core psychology.

The most obvious way to resist NP is to suppose that some admissible sequences are mixed: some of their determinates are mental while others are not. But this requires an additional principle explaining why certain natural

occupants of positions in the very architecture constitutive of core psychological theory fail to count as mental. In the absence of such a principle, the more natural conclusion is that the determinates occupying those positions are themselves mental properties. The mixed-sequence hypothesis is especially unstable because the allegedly non-mental determinates would not merely occur alongside mental properties. They would occupy positions in the same psychological architecture and enter into the relations that are characteristic of the standard mental roles. If such properties were nevertheless non-mental, it would be unclear why $\mathcal{P}$ should be regarded as characterizing the psychological domain in the first place.

The argument is strongest when we distinguish the case for mentality from resemblance to familiar mental kinds. Familiar phenomenal substitutions provide an easy entry point, but they are not the ultimate basis for NP. Suppose, first, that an unfamiliar phenomenal character occupies the determinable position occupied by pain in us. Nagel's (1974) reflections on bat experience provide the familiar reason for taking seriously phenomenal characters outside our own repertoire. Once familiar phenomenal rearrangements are admitted, the mere fact that a phenomenal property is absent from human psychology provides no principled reason to exclude it as an occupant of a psychological role.

More demanding cases involve non-phenomenal occupants. Let z -pain be a natural, *non-phenomenal* property proposed as an alternative determinate 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' (2019) view of the epistemic role of consciousness gives a principled version of this sort of objection. But what follows is that determinate substitutions can impose constraints on one another, not that non-phenomenal determinate sequences are impossible. Z -pain may require z -introspection, and that substitution may in turn require further coordinated substitutions. If a coordinated sequence preserves the relations specified by $\mathcal{P}$, including the relevant forms of intensional non-well-foundedness, then the determinable architecture has not changed. What has changed is the way its determinate occupants make those relations possible.

Eklund's recent discussion of alien minds suggests an even more unfamiliar possibility. He asks whether creatures might lack phenomenal consciousness while possessing some positive alternative to it, an alien feature playing in their mental lives something like the role phenomenal character plays in ours (Eklund 2026; see also Lee 2019). We need not assume that such a feature is possible. The dialectical point is that NP gives us a reason not to reject it merely because it is alien to us. If a natural sequence containing such properties genuinely instantiated the organization specified by $\mathcal{P}$, we would have the same structural reason for treating its occupants as mental that motivated NP in the first place. The grounds for mentality would lie in their participation in a natural, integrated psychological sequence rather than their resemblance to familiar mental properties.

The suggested failure of the Uniqueness Hypothesis has implications beyond undermining Bealer's non-reductive definitions. The alternative mental sequences give us an existence argument for mentality beyond the standard mental. Determinable-level functionalism thus moves us from multiple realizability to pluralism at the mental level 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 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 mentals *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 mentals and what is essential to mindedness in general. But NP undermines this support. The standard mentals 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 mentals 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 with which we are familiar. 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 thereby identified is the unique possible architecture of psychology.

Once we admit the possibility of nonstandard mentals, a further form of mental pluralism falls naturally into view:

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

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 mentals.

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, 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).

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. For, once it is granted the standard mental states themselves constitute only one kind of mind among others, AU becomes an independent restriction on the space of possible psychology and therefore 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 mental states. 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 another possibility: 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).

These cases need not all be understood as counterexamples to AU. Alief might belong to a more adequate account of the standard architecture, and an extended cognitive pathway might differ only in respects that $\mathcal{P}$ leaves unspecified. Their significance is that they make alternative forms of psychological organization intelligible.⁵ 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 $\mathcal{P}$ remains an open question. Whether a particular proposal changes a relation specified by $\mathcal{P}$, or only a feature that $\mathcal{P}$ leaves unspecified, remains an open question.

Functional characterization makes it possible to generalize this procedure. The roles specified by $\mathcal{P}$ are interdependent in that each is characterized partly through its relations to others. Suppose, for example, that $\mathcal{P}$ characterizes desire partly through its interaction with binary belief in practical reasoning. A credence-only psychology cannot simply remove belief while leaving that characterization untouched. It must specify how its motivational states interact with graded attitudes instead. Since $\mathcal{P}$ includes only constitutive relations, this substitution changes the desire-role itself. The consequences extend to any further roles whose definitions depend on

⁵ Shea (2020) develops a closely related anti-parochialist use of functionalism. He argues that functionalism need not require radically different creatures to instantiate the standard mental states, but can instead allow them to possess their own collection of functionally interrelated psychological states. Accordingly, he distinguishes the Ramsey sentences characterizing human psychology from a potentially different set characterizing Martian psychology. This provides support for the possibility of alternative psychological architectures contemplated here. The present argument differs in approaching that possibility through the failure of uniqueness within non-reductive functionalism and the resulting progression from narrow to broad mental pluralism.

the altered characterization, including intention, deliberation, and introspection. A revision introduced at one position therefore changes the architecture and may ramify throughout the interconnected definitions.

The result need not be a defective instance of $\mathcal{P}$. It may instead be an alternative psychological architecture, $\mathcal{P}_M$, specifying a coherent organization of mutually dependent properties. This differs from replacing the determinate occupants of fixed roles, as in narrow pluralism, since the revision changes the roles themselves. Wherever those changes ultimately reach, the resulting properties cannot be identical to the standard mentals whose constitutive roles have been altered. A sequence satisfying $\mathcal{P}_M$ may therefore consist of entirely alien properties.⁶

There is, moreover, no evident reason why such a revision would eliminate the 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 $\mathcal{P}$. These possibilities concern natural, irreducible properties at the psychological level, just as the alternatives considered under narrow pluralism do.

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 supply *prima facie* intelligible starting points, and the interdependence of functional definitions supplies a mechanism by which they can be comprehended as 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. Simply observing that it fails to instantiate $\mathcal{P}$ does not answer that reason.

Narrow pluralism does not entail broad pluralism. It does, however, undermine the identification of mentality with membership in the standard mental sequence. The alternative architectures considered here take the argument a step further insofar as they 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.

Without such an explanation, AU threatens to reproduce at the level of architecture the chauvinism functionalism was originally designed to avoid. Classical functionalism rejected substrate chauvinism. Similarly, narrow pluralism rejects a corresponding restriction to our determinate mental properties. AU nevertheless permits alien mental properties only if they reproduce the essential functional architecture abstracted from the standard mentals. Call this architectural chauvinism. Of course, our own psychology provides a methodological starting point

⁶ I leave open the possibility that the change may be consistent with local pockets of the standard mentals that are unaffected by the change.

for investigating the space of possible minds. But treating its constitutive architecture as a metaphysical boundary on mindedness is a much less defensible position.⁷

4. Philosophical implications

Thus far I have argued that philosophy of mind should take the possibility of alien minds seriously. The possibility of minds that differ, perhaps dramatically, from our own in both intrinsic qualities and functional design should be assumed as a working hypothesis. If so, this 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 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 mental states) fall. Even though the general metaphysical structure is familiar, these properties already involve a bit of conceptual engineering, broadly understood (Cappelen 2018). The failure of Bealer’s non-reductive definitions in light of non-standard sequences seemed to require building out new conceptual space to accommodate the class of possible minds contemplated by narrow pluralism. Our ordinary conception of the mind and mental properties doesn’t clearly distinguish (if it distinguishes it at all) a level of mental determinables of which the standard mental states are but one determinate. Narrow pluralism, in turn, raised in a particularly acute way the possibility of architectural chauvinism (and, hence, broad pluralism). And while imagining alien cognition is not foreign even in pop culture, it isn’t clear how exactly it is to be understood.⁸

These developments, however, merely set the stage for the real engineering work. Specifically, narrow pluralism doesn’t just give us the possibility of non-standard minds, but a range of non-standard mental properties that differ *intrinsically* from the standard mental states. Similarly, broad pluralism gives us an open-ended range of non-standard mental properties and relations that differ *functionally* from the standard mental states. Understanding the mental properties within this conceptual space even in general terms involves us in some of the most difficult problems in cognitive ethology writ large, and without the benefit of phylogenetic or ontogenetic constraints to guide us (Allen & Bekoff 1999, Bekoff, Allen, & Burghardt 2002).

Determining the features of even a part of this conceptual space will not be a simple matter, but we need to get at least a partial handle on it. The reason is that mental pluralism generates a conceptual-engineering problem that propagates beyond the philosophy of mind itself. At the first stage, we need concepts that adequately represent at least some of the ways in which possible minds might differ from our own. Without such resources, we cannot determine whether philosophical theories developed around the standard mental states genuinely extend to nonstandard psychologies. A theory of justification that appeals to phenomenal experience, for example, cannot simply be

⁷ 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).

⁸ Consider, for instance, the characters Spock (from *Star Trek*) and 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.

assumed to apply to a mind whose relevant experiential states are non-phenomenal (Moffett 2025; Butzer 2026) and the same goes for a theory of welfare built around phenomenal experience (Levy 2024). Similarly, a theory of agency built around a familiar belief-desire psychology cannot be assumed to apply to pushmi-pullyu or credence-only beings (Wong 2025; Collins 2026; for the animal case, see Steward 2009).

But establishing that such a theory fails to generalize creates a further problem. Suppose an alien system lacks the psychological features in terms of which our theory analyzes or grounds justification, agency, welfare, or some other philosophically significant property (cf. Block 1978). There are then three possibilities. First, the property may nevertheless apply to the alien system. In that case, we need to analyze or ground the property in a way that does not build its specifically human psychological realization into it. Second, we may deny that the property applies because the alien system is not, after all, genuinely a knower, agent, welfare patient, or other relevant subject. But such a response requires an independent reason for making possession of the familiar psychological features a condition on belonging to that more general category; otherwise, it simply relocates the original parochialism. Finally, we may accept that the system is genuinely intelligent or agential while denying that the target property applies to it. In that case, the property itself turns out to have a more restricted scope than our theorizing may have assumed. The resulting conceptual problem is then not how to extend justification, agency, or welfare to alien cases, but whether we require some more general category under which both the familiar and alien cases fall. For instance, if, as Smithies (2012, 2019) suggests, zombies cannot be justified in virtue of their lack of phenomenal properties, does this show that zombies are not subject to epistemic normative evaluation? Or does it show that justification is itself too parochial a concept to capture the relevant epistemic landscape?⁹

The general point is that mental pluralism forces us to make explicit distinctions that a focus on the home case allows us to leave obscure. Alien minds force us to determine which features of familiar cases are constitutive of the phenomena under investigation, which merely reflect one way those phenomena are psychologically realized, and which mark only local rather than genuinely general categories. The necessary conceptual work is therefore not simply a matter of broadening existing concepts to accommodate unfamiliar cases. It is the more fundamental task of locating the appropriate level of abstraction at which similarities and differences across possible minds should be represented. Alien minds matter philosophically because they deprive us of the assumption that the conceptual joints visible in the human case are the joints at which a general theory should be carved.

From the point of view of purely theoretical philosophy, these are important philosophical problems that make the methodology of philosophy considerably more difficult. For they show that even a fully adequate theory of human justification, human agency, or human welfare is, at most, a starting point for a general theory. And from that purely theoretical perspective, this might not be so troubling. After all, if we can answer these questions in a way that is extensionally satisfactory for humans, the effort to fully generalize the concepts can be treated as a downstream aspiration. Unfortunately, the rapid development of increasingly intelligent artificial systems has turned this theoretical exercise into an applied one. There is no guarantee that these systems won't differ from us in the

⁹ In the latter case, the conceptual engineer will then want to know whether the concept of justification itself needs to be adjusted.

ways allowed by both forms of mental pluralism.¹⁰ Addressing these issues matters for both AI alignment (e.g., see Josifović & Noller 2026, Moffett 2026) and AI welfare (Shepherd 2024, Goldstein & Kirk-Giannini 2025, Veit 2026). It is perhaps darkly ironic that philosophizing about zombies and aliens—paradigmatic cases of apparently impractical speculation—has become one of the most pressing practical issues of our age.¹¹

References

- Allen, C., & Bekoff, M. (1999). *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.
- 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. (2018). *Fixing Language: An Essay on Conceptual Engineering*. Oxford University Press.
- 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.
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19.
- Collins, S. (2026). Decision-making procedures explain group agency. *Inquiry*, 69(4), 1502–1521.
- 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.
- 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.
- Jackson, E. G. (2020). The relationship between belief and credence. *Philosophy Compass*, 15(6), e12668.

¹⁰ 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.

¹¹ It is also perhaps darkly ironic that the same practical urgency did not really gain traction around questions of animal welfare—perhaps because cows don’t factor into anyone’s p(doom) scenario.

- Josifović, S., & Noller, J. (2026). Agency and alignment: Toward a normative architecture for human–AI interaction. *AI & Society*, 41, 5561–5571.
- 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, 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.
- Millikan, R. G. (1995). Pushmi-pullyu representations. *Philosophical Perspectives*, 9, 185–200.
- Moffett, M. A. (2025). *The Indispensability of Intuitions*. Cambridge University Press.
- Moffett, M. A. (2026). Traversing the uncanny valley: Artificial wisdom and the limits of alignment. Unpublished manuscript.
- 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.
- 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). Functionalist 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.
- 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* (pp. 35–42). 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* (pp. 403–427). Kluwer Academic Publishers.
- 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.
- Steward, H. (2009). Animal agency. *Inquiry*, 52(3), 217–231.
- 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.